Doctor performance assessment

development and impact of a new system

To my parents

The studies presented in this thesis have been performed at the Scientific Institute for Quality of Healthcare (IQ healthcare). This institute is part of the Nijmegen Centre for Evidence Based Practice (NCEBP), one of the approved research institutes of the Radboud University Nijmegen and the Netherlands School of Primary Care research (CaRe), acknowledged by the Royal Dutch Academy of Science (KNAW).

The studies described in this thesis were supported by grants from 'De Orde van Medisch Specialisten' and the Netherlands Organisation for Health Research and Development (ZonMW).

Financial support by the Netherlands Organisation for Health Research and Development (ZonMW) and the Scientific Institute for Quality of Healthcare for publication of this thesis is gratefully acknowledged.

Nijmegen, 2011

Cover design: Jasper Visser
Lay-out: Jolanda van Haren/ Karlijn Overeem
Print: GVO drukkers en vormgevers B.V. | Ponsen & Looijen, Ede
ISBN: 978-90-817924-0-0

Doctor performance assessment

development and impact of a new system

een wetenschappelijke proeve
op het gebied van de Medische Wetenschappen

Proefschrift

Ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 15 november 2011
om 15.30 uur precies

door

Karlijn Overeem

geboren te Tilburg op 22 januari 1980

Promotor	prof. dr. R.P.T.M. Grol
Copromotoren	dr. H.C.H. Wollersheim mw. dr. M.J.M.H. Lombarts (AMC/ Universiteit van Amsterdam) dr. O.A. Arah (University of California, Los Angeles, Verenigde Staten)
Manuscriptcommissie	prof. dr. R.F.J.M. Laan mw. prof. dr. D.D.M. Braat prof. dr. C.P.M. van der Vleuten (Universiteit Maastricht)

CONTENTS

Chapter 1

Introduction

Concerns about the quality of health care and the performance of health care professionals persist.[1,2] Research shows that in (too) many cases suboptimal healthcare is delivered. This includes misuse, underuse and overuse of care.[3-5] Over the past decades, a variety of approaches have been suggested and tested to improve quality of health care. Focusing - on the one hand - on quality improvement through better availability and presentation of evidence to clinicians. For examples via feedback, reminders and educational meetings. The effect sizes of these strategies are however modest and no strategy appears to be consistently effective.[6] On the other hand, health care professionals and policy makers focused on organisational change and health care system performance. Several countries adopted organizational strategies such as disease management and clinical governance to overcome the quality and safety challenges of the 21st century. So far, it is unclear which organizational interventions result in higher quality and safer patient care.[7]

Additionally, in the last ten years, there is an increased emphasis on doctors' individual performance in improving the quality of healthcare.[7,8] There are many reasons for this, including societal developments such as the increased ability of patients to access detailed and accurate information about their own health and illnesses, the growing size of multidisciplinary teams and the increasing demand for evidence based practice. These societal developments require not only good medical knowledge, but also good communication, collaboration, organization and self-reflection skills. Nowadays, the consensus is that these competencies need to be assessed and supported continuously to ensure doctors perform optimally.[9] As a result, many countries are currently developing and implementing systems to assess doctors' performance. This prompts a lot of discussion and questions. This thesis aims to add to the understanding of doctor performance assessment in daily practice.

In this introductory chapter we will address four issues. In the first subsection doctor performance assessment is introduced and definitions will be provided. In the second paragraph we present an overview of doctor performance assessments in the United States (US), Canada and the United Kingdom (UK), where much of the terminology and experiences originate from. The section concludes that the current literature does not offer all the insights necessary for the development and implementation of acceptable and effective doctor performance assessments. In the third subsection we provide the context of this thesis and the Dutch health care system. The fourth and last subsection summarises the research questions and outline of the thesis.

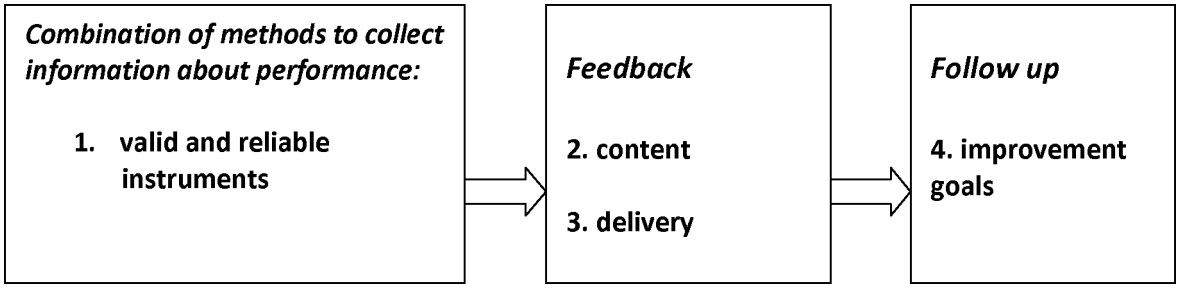
1. DEFINITIONS

Doctor performance can be defined as 'what a doctor actually does in practice' whereas competence means 'what a doctor is capable of.' In 1991, Rethans et al reported a discrepancy between how doctors perform in controlled examination situations and their behaviour in real practice.[10] Due to these observed differences, competency-based assessments can be defined as 'measures of what doctors do in testing situations', while performance-based assessments can be

defined as 'measures of what doctors do in practice'.[11] Generally, performance assessment can serve summative as well as formative goals. Summative assessments aim to support the decision whether a doctor is fit for practice whereas the primary goal of formative assessments is to give doctors insights into their performance and provide a direction for continuous professional development.

By a formative performance assessment system we actually mean: 1) instruments and methods to evaluate doctors' professional performance, 2) acceptable and effective feedback content and delivery and 3) follow up related to this feedback to individual doctors. These three components are presented in Figure 1. The first step includes a combination of methods to assess professional performance in a number of domains which encompass valid and reliable instruments. Epstein and Hundert defined professional competence as 'the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in daily practice for the benefit of the individual and community being served.[12] Literature on performance assessment shows that the incorporation of information from multiple sources and various occasions is essential in order to evaluate a complex construct such as doctor performance.[13] The second step implies that the information gathered in step one is being processed to the doctor involved, also called effective and acceptable feedback content and delivery. Several reviews have established broad agreement on characteristics of feedback content making it most effective.[14-16] Feedback should focus on task performance and should not contain any judgments about the character of the recipient. Furthermore, feedback should be clear and specific. Feedback delivery implies the way feedback is offered to the doctor assessed. This can be in a mailed feedback report or in an interactive manner, such as a discussion. The effects of delivering feedback in an interactive manner with a coach have been established in quality improvement research. Winkens et al found that personalized feedback, provided by a credible source such as a colleague can be effective in changing the quality and quantity of tests requests.[17] The third step of follow up is the use of the assessment for practice change and improvement. It includes improvement goals and the process of following up on them. The importance of follow up for the impact of feedback on performance has been highlighted by many researchers in the field of quality of care research. For the optimal effect of feedback, feedback is ideally part of a system of continuous monitoring including systematic evaluation of progress.[18] Thus, follow up is not the fact of improvement goals and whether they are achieved or not but the whole process after the feedback is proceeded. This can be an ongoing process of coaching or support focused at improvement in practice.

Figure 1: Elements and procedures of formative assessment system



2. A BRIEF DESCRIPTION OF DOCTOR PERFORMANCE ASSESSMENTS

United States

Doctors in the United States (US) have to demonstrate that they are fit to practice in a process called Maintenance of Certification (MOC). The authority for the performance of doctors has been delegated to accrediting organizations such as the American Board of Medical Specialists (ABMS) and the Joint Commission on the Accreditation of Healthcare Organizations. In the US, the government has, to date, played a secondary part in doctors' performance assessment. In the MOC program six competencies -identified in the ACGME (Accreditation Council for Graduate Medical Education) Outcome project- are assessed. These competences are: medical knowledge, patient care, practice based learning and improvement, interpersonal and communication skills, professionalism and systems based practice. The assessments encompass open-book tests, closed book tests, and self-assessments.[19] An instrument to measure humanistic qualities such as communication and collaboration was developed in 1993 by the American Board of Internal Medicine (ABIM).[20] Doctor P. Ramsey, internist, was the first to demonstrate that it is feasible to obtain reliable, multidimensional peer evaluations of individual doctors. This instrument is nowadays a voluntary part of the MOC program. Research showed that 65 percent of participants intend to change behaviour after receiving feedback from colleagues and patients on their performance.[21] An assessment of doctors' performance in actual practice with regard to competencies such as communication and collaboration is not an obligatory part of MOC yet. However, specialty societies, state medical boards and provider organizations or payers are evolving performance assessments in real practice.[8] Doctor performance assessments in the US are being affected by different authorities with contrasting interests and different efforts. Within the private (non-governmental) sector of the US, production and dissemination of quality-related data is driven by groups of large companies who have become advocates for public reporting, pay-for-performance and improved patient safety. Those companies press for more detailed information on both the quality and the cost of services provided to their employees by individual health plans. Healthcare insurers and health plans set up processes for measurement of doctors' performance to produce quality-related data.[22] The emphasis is on performance indicators relating to processes and outcomes or pseudo-outcome measurements (for example organizational aspects, HbA1C-levels and adherence to guidelines).[23] In addition, hospitals in the US must provide quality-related data to the Joint Commission on the Accreditation of Healthcare

Organizations. Those data collected by healthcare insurers and health plans claim to be aimed at informing value-based decisions by consumers of health care and not so much on doctor performance improvement. Indeed, it has been shown that doctors are often not responsive to publicly released information and these quality related data do not automatically incite performance improvement.[24-27] Moreover, it has been highlighted that care should be taken in assessing doctors based on narrow performance measures.[22] Next to this, the sue culture and liability risks in the United States affects performance assessments. Kesselheim and Donohue point to the fact that physicians in the US might be reluctant to embrace doctor performance assessment initiatives on the grounds that they will be used as evidence against doctors in malpractice litigation.[28]

Canada

In the year 2000, Canada started to describe the standards of competence, care and conduct expected of all doctors in the CanMEDS competences.[29] Subsequently, the assessment of those competencies in clinical practice started. In Canada, a major role for assessing doctors' performance is established by regulatory authorities and specialty societies. Each province has its own regulatory body -called College of Physicians and Surgeons- which is legislated to monitor doctor performance.[30] Since 1972, the College of Physicians and Surgeons of Ontario assessed doctors every five years with a so-called clinical audit. All doctors who turn 70 years of age in a given year are automatically selected for assessment, and the program assesses a random selection of doctors within specific practice and specialty areas. Assessments consisted of a tour of the practices and a review of medical records to evaluate the system of record keeping and the content of the records and thereby indicate the quality of a doctors' performance.[31,32] As these assessments appeared to be resource intensive, a more pragmatic approach was developed in 1996. The Physician Achievement Review program developed and standardized Multi Source Feedback (MSF) questionnaires for hospital doctors and family physicians. MSF gathers information from persons who are qualified and have credibility to judge clinical practice, such as 1) peers familiar with a similar domain of practice; 2) members of the health care team; 3) patients as the recipients of health care. Thus, MSF covers a wide range of perspectives and competencies based on observation. Nowadays, MSF is being used in several provinces for surgeons, paediatricians, anaesthetists, radiologists, family physicians, pathologists and international medical graduates.[33-36] Physicians are required to participate in those MSF assessments every 5 years. Once the questionnaires have been completed, the resulting feedback reports are reviewed by members of the Physician Performance Committee (PPC). Doctors in the lower percentile are being supported by the PPC and other competency assessment tools are used to measure their performance. All participating physicians receive a mailed MSF report. Research demonstrates that 66 percent of physicians report having initiated a change for at least one aspect of practice as a result of the MSF report.[37] However, it is not yet clear whether doctors succeed in implementing and maintaining a change. A 5 year longitudinal

study found that only small changes in performance occur.[38] Focus group studies amongst family physicians in Canada revealed that feedback is only useful if it is perceived to be accurate and credible; feedback perceived as negative and inaccurate is less likely to lead to practice improvement. In addition, interviews revealed that feedback was often perceived as not specific enough to unravel needs for improvement. In addition, it became clear that often emotional reactions occur when negative feedback is delivered without facilitation.[39]

United Kingdom

The General Medical Council (GMC) in the United Kingdom (UK) regulates British doctors through the Medical Act. The Council comprises doctors (who predominate) and laypeople. It registers doctors for UK practice, sets professional standards, regulates basic medical education, and manages doctors' fitness to practice.[40] In the United Kingdom, employment based assessments predominate.

After a series of medical scandals including the 1990s crisis in paediatric cardiac surgery at Bristol Royal Infirmary, public pressure accelerated radical change. The General Medical Council (GMC) introduced a system of 'revalidation' -the process in which doctors prove they are up to date and perform up to standard- in 1998 as a way to win back the trust of the British public.[41] Since then, doctors in the UK are expected to demonstrate that they remain up to date and fit to practice themselves. Revalidation started with the requisite that doctors maintain a folder which contains information about how they practice. This can include: certificates of training, results of significant event analysis, audits, patient satisfaction surveys and complaints. The primary purpose of revalidation is summative, which means that the outcome is to decide upon a doctors' certification.

These centrally driven initiatives have been made politically possible because of a substantial number of high profile cases in the UK, in which quality of care has been a serious problem. Although associated with a substantial loss of autonomy among doctors, this initiative includes a regular assessment of individual doctors. Alongside this, annual appraisals were introduced. Appraisals were set up as an opportunity to plan improvement and equip doctors for lifelong learning. It consisted of a formative interview with the aim of 'facilitated self-reflection' with a trained colleague, called appraiser.[42] However, by April 2003, the GMC decided that revalidation would be based on doctors' annual appraisal forms.[43] This meant that appraisal was used for summative goals as well as formative goals. This is considered as an undesirable and inconvenient development. Together with the introduction of revalidation and appraisal, there was a need for reliable methods of assessing doctors' competence and performance. As a result, the GMC started in 2004 with the development and introduction of questionnaires completed by patients and colleagues just like in Canada as a means of obtaining multisource feedback on the performance of individual doctors.[44] Colthart et al conducted a survey study amongst Scottish general practitioners (GPs) over 3 years to examine their opinions on the relevance and impact of appraisal. Thirty-three percent of responders reported undertaking further education or training

as a result of appraisal, and 13 percent felt that appraisal had influenced their career development. However, many doctors -54 percent- perceived limited benefit.[45] This is in agreement with other studies.[46,47] Especially the link of appraisal with revalidation (summative aims) is found to be problematic.[48] Since the year 2008 MSF is an obligatory part of the workplace based assessment of residents and it has been introduced for GPs in revalidation. Junior doctors in the UK perceived low effectiveness of MSF.[49] They believed the MSF tools were unable to effectively identify doctors in difficulty or provide developmental feedback. Amongst GPs in the UK it was found that MSF on clinical behaviour was not perceived to be useful.[50]

3. RATIONALE UNDERLYING THIS THESIS

Irrespective of national borders, a look at today's landscape of doctor performance assessments, prompts the conclusion that the growing demands for greater public accountability have propelled regulatory bodies towards the introduction of a noteworthy set of performance assessment and certification tools. The emphasis globally is on summative goals and legislation aspects. One of the methods most frequently used in the assessment of doctors' performance is MSF. However, research into the formative potential of MSF (in other words its impact on future performance) is still in its infancy. Studies in Canada and the United Kingdom highlighted that the educational impact of MSF might be limited by the content and delivery.[51,52] Interviews revealed that due to emotional reactions and lack of specificity, the acceptance of feedback was hindered. As a result, it has been advocated that the delivery of MSF should include a formal mentor or coach to increase the acceptance of feedback (step two and three of Figure 1). However, this has not been investigated and tested in practice yet. In sum, despite the growing literature on doctor performance assessment, it is still not clear how we move from valid and reliable MSF instruments to systems with 1) a combination of feasible methods, 2) acceptable and effective feedback content and delivery, 3) sustainable impact and transferability across settings and doctors.

4. RESEARCH CONTEXT

The Dutch health care system can be characterised as a regulated market oriented system within the context of a universal insurance system.[52] Hospitals and doctors –legally- share the responsibility for quality of care whereas doctors are considered a self-regulating profession. The majority (approx. 70 percent) of the medical specialists (approximately 16.000) in the Netherlands are independent entrepreneurs, per specialty organised in 'partnerships', who are paid through a fee-for-service system. The other part of the medical specialists are employed by a general or academic hospital. At hospital level, specialists are organised in a medical staff through which they participate in hospital management. Health care professionals should develop their own quality assurance mechanisms in contributing to a transparent health system. For the medical profession this development took place, for a great part, within the framework of the specialty societies. The functioning of a group of specialists in a partnership is evaluated by

'visitatie', which was introduced in 1989.[53] It is a program for external peer review through site-visits. The program is a doctor-led and -owned quality assurance activity, meaning that doctors set the standards, conduct the surveys, formulate the recommendations for improvement and decide upon corrective actions.[54]

In 2005, the Royal Dutch Medical Association (KNMG) -the umbrella organization for all medical doctors (including general practitioners and doctors in social medicine) published a policy document which states that all qualified doctors should be evaluated on a regular basis. [55] Moreover, in the Netherlands there has been a shift in postgraduate medical education since 2005. Trainee doctors are required to record evidence of their competence and take an active role in their own development.[56] Since the performance of individual doctors is not the primary focus of visitaties, a new program had to be developed, implemented and studied.
As a logical consequence, the Central Board of the Dutch Specialists Organization (OMS), established an expert panel in June 2005 with the intention of introducing a peer led performance assessment system for individual medical specialists, called IFMS (Individueel Functioneren Medisch Specialisten).[57] This is the subject of this thesis.

Research questions

Although the need for doctor performance assessment is clear, uncertainty remains about the optimal methods and design of doctor performance assessment systems. Given the importance of the assessments made, in terms of both patient safety and a doctor's personal development, it is essential to develop and evaluate assessment programs thoroughly. The main argument of this thesis is to provide a performance assessment system composed of effective and feasible methods and reliable and valid instruments to assess and improve the professional performance of medical specialists in the Netherlands. The research questions are classified according to the three main themes of this thesis: methods and instruments used, the optimal design of a formative performance assessment system and the contextual factors influencing the educational impact.

This thesis addresses the following research questions:

Methods and instruments to assess doctors' performance

1. What are the psychometric properties of existing instruments and the feasibility and impact of methods currently available for the assessment of doctors' performance?
2. What are the psychometric properties of three new MSF instruments used by colleagues, coworkers and patients to evaluate a doctors' performance?

Design

3. What are the feasibility and perceived educational impact and topics addressed of a newly developed performance assessment system combining MSF with portfolio learning and a mentor?

4. How do mentors perceive and fulfil their role in doctor performance assessments?

Contextual factors

5. Which factors are incentives, or disincentives, for specialists to implement suggestions for improvement from MSF?
6. Which factors influence the reported change as a result of MSF amongst medical specialists in the Netherlands most?

Thesis outline (see Table 2)

Chapter 2 systematically evaluates the psychometric properties of existing instruments and the feasibility and effectiveness of methods for the performance assessment of doctors. We have seen in this introduction that many countries implement MSF. For a better understanding and insight in the field we performed a systematic review to explore other methods to assess doctors' professional performance. The aim is: 1) to investigate the validity and reliability of instruments used to assess professional performance; 2) to evaluate the feasibility and effectiveness of the different methods used.

Chapter 3 continues by implementing the performance assessment system with assessment instruments into a larger cohort of 26 hospitals, including 146 doctors that were assessed. The aim of the study was to investigate the reliability and validity of the new instruments used for assessment by colleagues, coworkers and patients. The study design was an iterative, developmental validation study of three MSF instruments.

Chapter 4 kicks off with the implementation of a performance assessment system for medical specialists in the Netherlands. The aim of this study is to investigate the feasibility, topics addressed and impact of performance assessment for medical specialists in eight self-selected hospitals in the Netherlands. The study compares three methods of MSF. Data were primarily collected through semi-structured telephone interviews and a postal survey for mentors and doctors involved.

Chapter 5 focuses on the impact of performance assessments with MSF on future performance. Its aim is to evaluate the impact of MSF by using an adapted model for change in professional performance. This model identifies four steps for changes in professional performance: awareness of improvement needs, acceptance of improvement goals, taking actions and maintenance of change. The study illuminates characteristics of assessment systems that might explain performance improvement as well as other factors that determine the improvement of doctors' performance as a result of MSF. Data were primarily collected through semi-structured face-to-face interviews.

Chapter 6 explores how mentors perceive and actually fulfil their role in order to disclose elements of effective strategies for delivering feedback of external assessments and discussing the portfolio. A mentor can help doctors interpret the (multisource) feedback and critically analyze their performance making use of the feedback to guide future performance. However, it is not yet clear what strategies mentors actually use to make doctors aware of their performance

and encourage performance improvement. Data were gathered by 2 surveys and semi-structured face-to-face interviews with mentors.

Chapter 7 features an evaluation of the implementation in 26 hospitals. The aim of the study is to examine the acceptance and perceived impact of MSF and to quantitatively measure influencing factors on its use. The study consist of a quantitative evaluation including 246 specialists using regression analyses techniques.

In the final chapter 8, the general discussion, the results of all studies will be synthesised, strengths and weaknesses of the studies will be discussed and some implications for future research and practice will be drawn. On the basis of the studies included in the thesis we introduce a model for doctor performance assessments.

Table 2. Outline of the thesis

Chapter	Research aims	Theme	Design
Chapter 2	To systematically evaluate the psychometric properties of existing instruments and the feasibility and effectiveness of current methods of performance assessment	Methods and instruments	Systematic review
Chapter 3	To develop MSF instruments for medical specialists in the Netherlands and to evaluate its psychometric properties	Methods and instruments	An iterative development and validation study of three MSF instruments
Chapter 4	To implement a performance assessment system for medical specialists and to evaluate its feasibility, topics addressed and perceived impact	Design	Process evaluation based on quantitative methods
Chapter 5	To explore hampering and stimulating factors that determine the improvement of doctors' performance as a result of MSF	Contextual factors	Qualitative study based on semi-structured face-to-face interviews
Chapter 6	To explore how mentors perceive and fulfil their role in order to disclose elements of effective strategies for feedback delivery and encouraging reflection	Design	Mixed method design comprising 2 surveys and semi-structured face-to-face interviews
Chapter 7	To implement a performance assessment system nationwide for medical specialists in the Netherlands and to measure influencing factors upon its impact	Contextual factors	Quantitative study based on regression analyses

References

1 Grol R, Berwick DM, Wensing M. On the trail of quality and safety in health care. BMJ 2008;336:74-6.
2 Shine KI. Health care quality and how to achieve it. Acad Med 2002;77:91-9.
3 Mangione-Smith R, DeCristofaro AH, Setodji CM, et al. The quality of ambulatory care delivered to children in the United States. N Engl J Med 2007;357:1515-23.
4 Chassin MR, Galvin RW. The urgent need to improve health care quality. Institute of Medicine National Roundtable on Health Care Quality. JAMA 1998;280:1000-5.
5 McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003;348:2635-45.
6 Grimshaw J, McAuley LM, Bero LA, et al. Systematic reviews of the effectiveness of quality improvement strategies and programmes. Qual Saf Health Care 2003;12:298-303.
7 Grol R. Improving the quality of medical care - Building bridges among professional pride, payer profit, and patient satisfaction. JAMA 2001;286:2578-85.
8 Klass D. Assessing doctors at work--progress and challenges. N Engl J Med 2007;356:414-5.
9 Klass D. A performance-based conception of competence is changing the regulation of physicians' professional behavior. Acad Med 2007;82:529-35.
10 Rethans JJ, Sturmans F, Drop R, van der Vleuten CPM, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. BMJ 1991;303:1377-80.
11 Rethans JJ, Norcini JJ, Baron-Maldonado M, et al. The relationship between competence and performance: implications for assessing practice performance. Med Educ 2002;36:901-9.
12 Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287:226-35.
13 van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309-17.
14 Shute VJ. Focus on formative feedback. Rev Educ Res 2008;78:153-89.
15 Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007;77:81-112.
16 Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119:254-84.
17 Winkens RA, Pop P, Bugter-Maessen AM, et al. Randomised controlled trial of routine individual feedback to improve rationality and reduce numbers of test requests. Lancet 1995;345:498-502.
18 Grol R, Wensing M. Implementation. Effective change in patient care (in Dutch). Maarssen: Elsevier, 2001:243-60.
19 The American Board of Medical Specialties. Maintenance of Certification. www.abms.org/maintenance_of_certification. Accessed on 25-3-2011.
20 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655-60.
21 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med 2002;77:S64-S66.
22 Parkerton PH, Smith DG, Belin TR, Feldbau GA. Physician performance assessment: nonequivalence of primary care measures. Med Care 2003;41:1034-47.
23 Lanier DC, Roland M, Burstin H, Knottnerus JA. Doctor performance and public accountability. Lancet 2003;362:1404-8.
24 Werner RM, Asch DA. The unintended consequences of publicly reporting quality information. JAMA 2005;293:1239-44.
25 Donohue SK. Health care quality information liability & privilege. Ann Health Law 2002;11:147-58.
26 Faber M, Bosch M, Wollersheim H, Leatherman S, Grol R. Public reporting in health care: how do consumers use quality-of-care information? A systematic review. Med Care 2009;47:1-8.
27 Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med 2008;148:111-23.
28 Kesselheim AS, Ferris TG, Studdert DM. Will physician-level measures of clinical performance be used in medical malpractice litigation? JAMA 2006;295:1831-4.
29 Frank JR, Jabbour M, Tugwell P, et al. CanMEDS 2000:extract from the CanMEDS 2000 project societal needs working group report. Med Teach 2000;22:549-54.
30 Dauphinee WD. Revalidation of doctors in Canada. BMJ 1999;319:1188-90.

31 McAuley RG, Henderson HW. Results of the peer assessment program of the College of Physicians and Surgeons of Ontario. CMAJ 1984;131:557-61.

32 McAuley RG, Paul WM, Morrison GH, Beckett RF, Goldsmith CH. Five-year results of the peer assessment program of the College of Physicians and Surgeons of Ontario. CMAJ 1990;143:1193-9.

33 Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. CMAJ 1999;161:52-7.

34 Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. Can J Anaesth 2006;53:33-9.

35 Violato C, Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. Pediatrics 2006;117:796-802.

36 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. BMJ 2003;326:546-8.

37 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. Acad Med 1999;74:702-14.

38 Violato C, Lockyer JM, Fidler H. Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. Med Educ 2008;42:1007-13.

39 Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. Med Educ 2005;39:497-504.

40 Irvine D. A short history of the General Medical Council. Med Educ 2006;40:202-11.

41 Norcini JJ. Where next with revalidation? BMJ 2005;330:1458-9.

42 Conlon M. Appraisal: the catalyst of personal development. BMJ 2003;327:389-91.

43 Kmietowicz Z. Chief medical officer's report on the nation's health: poorly performing doctors are to be dealt with more fairly. BMJ 2003;327:69.

44 Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. Qual Saf Health Care 2008;17:187-93.

45 Colthart I, Cameron N, McKinstry B, Blaney D. What do doctors really think about the relevance and impact of GP appraisal 3 years on? A survey of Scottish GPs. Br J Gen Pract 2008;58:82-7.

46 McKinstry B, Peacock H, Shaw J. GP experiences of partner and external peer appraisal: a qualitative study. Br J Gen Pract 2005;55:539-43.

47 Lewis M, Elwyn G, Wood F. Appraisal of family doctors: an evaluation study. Br J Gen Pract 2003;53:454-60.

48 Boylan O, Bradley T, McKnight A. GP perceptions of appraisal: professional development, performance management, or both? Br J Gen Pract 2005;55:544-5.

49 Burford B, Illing J, Kergon C, Morrow G, Livingston M. User perceptions of multi-source feedback tools for junior doctors. Med Educ 2010;44:165-76.

50 Owens J. Is multi-source feedback (MSF) seen as a useful educational tool in primary care? A qualitative study. Educ Prim Care 2010;21:180-5.

51 Sargeant J, Mann K, Sinclair D, et al. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract 2008;13:275-88.

52 Seddon N. Is the future Dutch? Lancet 2008;372:103-4.

53 Lombarts MJMH, Klazinga NS. A policy analysis of the introduction and dissemination of external peer review (visitatie) as a means of professional self-regulation amongst medical specialists in The Netherlands in the period 1985-2000. Health Policy 2001;58:191-213.

54 Lombarts MJMH, Klazinga NS. Supporting Dutch medical specialists with the implementation of visitatie recommendations: a descriptive evaluation of a 2-year project. Int J Qual Health Care 2003;15:119-29.

55 Koninklijke Maatschappij ter Bevordering van de Geneeskunst. Het functioneren van de individuele arts. [Performance of the individual doctor] Utrecht: KNMG, 2005. (in Dutch)

56 Bleker OP, Hoorntje JCA, Schelfhout VJ. Beter en leuker, CCMS ontvouwt plannen voor de vervolgopleiding van medisch specialisten. Med Contact (in Dutch) 2004;59:1692-5.

57 Orde van Medisch Specialisten. www.orde.artsennet.nl/Nieuws/Artikel/Historie-IFMS. Accessed on 25-3-2011.

Chapter 2

Doctor performance assessment in daily practice: does it help doctors or not? A systematic review

Karlijn Overeem
Marjan J. Faber
Onyebuchi A. Arah
Glyn Elwyn
Kiki M.J.M.H. Lombarts
Hub C. Wollersheim
Richard P.T.M. Grol

Medical Education 2007:41:1039–1049.

Abstract

Context

Continuous assessment of individual performance of doctors is crucial for life-long learning and quality of care. Policy-makers and health educators should have good insights into the strengths and weaknesses of the methods available. The aim of this study was to systematically evaluate the feasibility of methods, the psychometric properties of instruments that are especially important for summative assessments, and the effectiveness of methods serving formative assessments used in routine practice to assess the performance of individual doctors.

Methods

We searched the MEDLINE (1966-January 2006), PsychINFO (1972-January 2006), CINAHL (1982-January 2006), EMBASE (1980-January 2006) and Cochrane (1966-2006) databases for English language articles, and supplemented this with a hand-search of reference lists of relevant studies and bibliographies of review articles. Studies that aimed to assess the performance of individual doctors in routine practice were included. Two reviewers independently abstracted data regarding study design, setting and findings related to reliability, validity, feasibility and effectiveness using a standard data abstraction form.

Results

A total of 64 articles met our inclusion criteria. We observed 6 different methods of evaluating performance: simulated patients; video observation; direct observation; peer assessment; audit of medical records, and portfolio or appraisal. Peer assessment is the most feasible method in terms of costs and time. Little psychometric assessment of the instruments has been undertaken so far. Effectiveness of formative assessments is poorly studied. All systems but 2 rely on a single method to assess performance.

Discussion

There is substantial potential to assess performance of doctors in routine practice. The longterm impact and effectiveness of formative performance assessments on education and quality of care remains hardly known. Future research designs need to pay special attention to unmasking effectiveness in terms of performance improvement.

Introduction

Whereas in the last decades the focus in improving quality of care has been on organisational change, we now see a gradual switch to include the assessment of the individual doctor's performance in day-to-day clinical practice.[1,2] This is a logical development: in 1927 Francis W Peabody, a prominent Boston physician, noted that 'the essence of the practice of medicine is that it is an intensely personal matter'.[3] Thus, in seeking to improve the quality of care, we need to focus on its central actor: the doctor. For several decades, initial certification was considered sufficient to guarantee quality for the entire professional life of a doctor. However, as medicine changes quickly and knowledge becomes outdated very fast, the consensus nowadays is that doctors need to maintain and develop their competences continuously.[4,5] A systematic review by Choudhry et al showed an inverse relationship between years in practice and the quality of care provided by a doctor.[6] Current investments in the education of medical students and continuous professional development for doctors are not enough to ensure that doctors perform optimally in their daily work.[7] Thus, it seems reasonable that doctors are supported in everyday practice if and when needed. As a consequence, performance assessment systems are being implemented worldwide. Performance assessment can serve 2 purposes: it can be either summative or formative. The former may support decisions for recertification or for remediation for underperforming doctors. The latter gives doctors insights into gaps in their knowledge, skills and competences and provides a direction for continuous professional development. The utility of summative as well as formative assessments is determined by their feasibility, reliability, validity and effectiveness.[8] These 2 types of assessment are linked to different goals: the most important criteria for summative assessments are validity and reliability, whereas formative assessments should be especially effective in improving performance.

Although the need for regular performance assessment of individual doctors is clear, the best way to do it is not. Current performance assessment systems emphasise competence evaluation.[9–11] Methods to assess professional competence have been investigated in detail.[12] However, research has shown a discrepancy between how doctors perform in controlled examination situations and their behaviour in real practice.[13] Different methods to assess doctors' performance in real practice are recommended in the literature and include peer assessment, the use of simulated patients (SPs) and video observation.[14–16]
However, their psychometric properties and effectiveness are not very clear.[17,18] It is important that policy-makers and educators in health care who are responsible for setting up systems for the assessment of individual doctors have good insights into the strengths and weaknesses of the different methods available.

The aim of this study was to systematically evaluate the feasibility of methods, the psychometric properties of instruments that are used in summative assessments, and the effectiveness of

methods used to deliver formative assessments in routine practice to assess the performance of individual doctors.

Methods

Inclusion of studies

Our strategy was based on a design for reviews in educational research.[19] We used different databases to search for articles that studied the reliability or validity of instruments or the feasibility or effectiveness of methods used for the performance assessment of individual doctors in routine practice. We performed searches in the following databases: MEDLINE, 1966-January 2006; PsychINFO, 1972-January 2006; CINAHL, 1982-January 2006; EMBASE, 1980-January 2006, and Cochrane, 1966-2006. All searches were limited to English language publications. We used the following National Library of Medicine medical subject headings: Clinical Competence (MeSH) OR Employee Performance Appraisal (MeSH) AND Methods (MeSH) AND Standards (MeSH) AND Physicians (MeSH). In addition, we searched reference lists of relevant studies and bibliographies of review articles. We also contacted authors of key references for additional information. The complete search is available from the authors.

Eligibility criteria

We included studies aimed at assessing individual doctors' performance in routine practice through the introduction of methods with or without a particular instrument. Because of the complex nature of educational and organisational interventions, it is not always appropriate to study feasibility and effectiveness using a randomised, controlled design and therefore non-randomised designs are often used.[20]

Subsequently, we did not make a selection according to the design of the study. Our inclusion criteria were broad. All studies conducted with general practitioners (GPs), hospital-based specialists or residents working in solo practices, group practices or hospitals were included. Moreover, studies had to offer psychometric data or data regarding the feasibility or effectiveness of methods to be included. Studies that measured competence in examination settings and studies concerning the performance of medical students, nurses or other health care professionals were excluded. Given that a pre-registration house officer is defined as 'a probationer doctor who still requires training and supervision', we included studies concerning junior doctors.[21] Patients are the end-users of health care and should therefore not be dismissed when evaluating routine practice of doctors. However, variables other than doctor performance, for example, patient demographics and health care setting, have been shown to influence patient satisfaction surveys.[22,23] We considered this to be a specialist topic which deserved attention in separate reviews.[24] We therefore excluded studies using patient-based assessment tools only.

Data extraction

Two reviewers (KO and MJF) independently made eligibility judgments based on article titles and abstracts. Discrepancies were resolved by discussion. We anticipated that studies would be too heterogeneous in design to be combined using a formal meta-analysis or to allow for quantitative analysis of data. A performance assessment system reflects the implementation of different methods with or without instruments to assess doctors, combined with proper procedures for processing the results and offering feedback to doctors with a specific purpose.[13] We defined a method as a way to collect information about an individual doctor. Within a method, various instruments can be used to produce quantitative or qualitative information. Two reviewers undertook data abstraction.

Data were extracted blinded onto a standard data abstraction form covering:
1 country of origin;
2 study population (primary or secondary care);
3 number of doctors included in the study;
4 study design;
5 information about the instrument concerning nature of scales and number of items;
6 validity of assessment;
7 reliability of assessment;
8 feasibility, and
9 effectiveness.

Feasibility was evaluated in terms of time and costs. We analysed effectiveness of formative assessment systems using a modified version of Kirkpatrick's model introduced by Curran and Fleet.[25] In this model, 4 levels of effectiveness are identified, namely: learner satisfaction; learning outcomes; performance improvement, and patient or health outcomes.

Assessment of study quality

A quality analysis was performed according to a strategy developed by an organisation for evidence in medical education.[26] We considered that the included studies differed too much in terms of methodology to compare reliability and validity data qualitatively. We performed a quality analysis for studies that investigated the effectiveness of assessment and feedback. We used numeric scales of 1-5 to assess study quality. Quality was independently assessed by 2 of the authors (KO and MJF). Any disagreement in quality scores was resolved by discussion.

Results

Search results

The search yielded 1184 articles (MEDLINE 366, PsychINFO 298, Cochrane 134, CINAHL 15, EMBASE 371). Studies not meeting our inclusion criteria (n=1140) were excluded. After reading

titles and abstracts, 44 articles were considered to be relevant. We collected 38 additional articles through manual searching of articles' bibliographies. Full papers were retrieved for 82 articles for detailed investigation. Subsequently, 18 articles were excluded because they did not meet our inclusion criteria. A total of 64 articles met the inclusion criteria, representing 58 different studies. These were predominantly uncontrolled, prospective studies with qualitative or quantitative evaluations. Most studies had been conducted in the UK, Canada, the USA or the Netherlands, among family doctors, hospital-based specialists and/or registrars.

We observed 6 different methods: SPs were used in 5 studies;[27–31] video observation was used in 9 studies; [32–40] peer assessment was used in 23 studies;[41–63] portfolio or appraisal were used in 11 studies;[64–74] direct observation was used in 3 studies,[75–77] and audits of medical records or written correspondence were performed in 10 studies.[78–87] Three studies used a combination of methods.[88–90] In 55 studies an explicit instrument to rate performance was applied.
A more detailed and structured summary of setting, domains, nature of the scales used and the number of items of instruments, psychometric properties, feasibility and effectiveness is given in tabular form on our website (http://www.wokresearch.nl) and is available from the authors.

Types of assessment methods and/or instruments
Methods can be either direct or indirect evaluations of performance. Direct methods concern observations of actual doctor-patient encounters and indirect methods retrospectively reflect the result of a doctor-patient interaction. Simulated patients, video observation and direct observation are direct methods, whereas peer assessment, portfolios or appraisals and audit of medical records are examples of indirect evaluations.

1 Simulated patients. Five studies investigated the use of covert SPs in routine practice. An incognito SP visits a doctor and rates his or her performance using pre-defined criteria. Ratings by SPs in 1 study were checked with an expert panel, which rated the tape-recorded consultations.[30]
2 Video observation. Nine studies carried out in the Netherlands and the UK explored video observation in the routine practice of GPs. Doctor consultations were videotaped and scored by 1 or 2 observers.
3 Peer assessment. In 23 studies peer ratings were used to provide an indicator of a doctor's performance. Medical colleagues or coworkers completed confidential questionnaires regarding knowledge, communication skills, professionalism, management and collegiality. Sometimes, patient ratings were added. Other terms used instead of 'peer assessment' are '360-degree feedback' and 'multi-source feedback'.
4 Portfolio or appraisal. Portfolio or appraisal were investigated in 11 studies. Appraisal refers to a structured process of facilitated self-reflection.[69] A portfolio or appraisal folder may be described as a collection of evidence maintained and presented for a specific purpose.[65] A mentoring system is a key element for portfolios and appraisals. The UK introduced appraisals

in April 2005 in general practice. Portfolio is being studied for paediatric registrars and GPs in the UK.[65,68]

5 Direct observation. In 3 studies doctors were observed and assessed directly during consultations, ward rounds or in the operating theatre. In all studies, an instrument was used to rate performance on, for example, history taking, physical examination and/or communication skills.[75–77]

6 Audit of medical records. Audit of medical records or referral letters can be used to measure doctor performance, as reported in 10 studies. For instance, Norton et al developed the Peer Assessment Program in Canada in 1984 to assess medical records.[85] Doctors being assessed in this programme in Canada were randomly chosen from the College's register, were above the age of 70 years or were directly referred by a complaint committee.[85]

Feasibility of methods in routine practice

The different methods differ considerably in feasibility in terms of time and costs (Table 1). The estimates of time and costs reported are based on what is required to achieve reliable results. This is not the case for portfolio or appraisal because these data were not available in the included studies. Peer assessment is most convenient in terms of time, with an average time investment of 1 hour per doctor.[63] Portfolios and appraisals are most time-intensive for doctors. The preparation of an appraisal folder or portfolio takes between 15 and 40 hours per doctor.[69,70] Peer assessment is most affordable and was estimated as costing £107 (€158) per doctor.[41] Simulated patients, portfolio or appraisal and video observation are most expensive. The calculated cost per completed appraisal was £771 (€1135).[70] Video observation was estimated as costing £268 (€394) per doctor and assessment with SPs, which requires at least 6 SP visits, as costing between £3 (€4) and £142 (€209) per visit, depending on the type of case presented.[27,37]

Utility of methods as summative assessment

RELIABILITY

Reliability of assessment instruments concerns internal consistency and stability (inter-rater reliability, intra-rater reliability or generalisability).[91] Generally, the instruments applied appeared to have reasonable internal consistency, with Cronbach's alphas varying from 0.83-0.98, except for 2 studies.[54,77] A couple of studies report inter-rater reliability. However, generalisability offers a better insight into reliability because it takes into account different sources of variance, such as variance of the cases, variance of the rater and interaction between the case and the rater.[92] The generally accepted threshold of reliability for high stakes judgement is a generalisability coefficient of 0.8.[92] Generalisability is established for all methods except for direct observation. As can be seen in Table 1, the number of raters needed ranges from 5 to 11 for methods in which 1 case is assessed. Seven to 11 raters are necessary for peer assessment,[42,52] and 5 raters for portfolios or appraisal folders.[68]

In SP assessment, video observation and medical audit, it is important to take into account the variance of the case presented when analysing generalisability. The number of cases and raters needed to achieve reliable results are comparable for SPs and video observation. Reliable results are achieved with 14 cases for SPs[28] and 12 cases for video observation.[37,39] For medical audit we looked at a combination of cases and raters. The numbers needed range from 5 raters judging 10 cases when referral letters are assessed,[81] to 3 raters assessing 60 cases that concern radiologist reports.[79]

Table 1. Overview of most important findings concerning validity, reliability and feasibility

	Validity	**Reliability**	**Feasibility (time and costs per doctor)**
Simulated patients	Content validity: Detection rate of simulated patients<8%[27,30] Criterion validity: not established[29]	G>0.8 is achieved with 6 cases for norm-referenced interpretation and 14 cases for absolute interpretation of scores[28]	Time: 3-7 hours of testing time, except time for preparation[27] Costs: depends on the case presented; £3- £39 (€4-€51) for lab tests and £14-£142 (€26-209) for imaging tests[27]
Video observation	Content validity of instruments is confirmed.[34-36] Tapes used are representative sample and consulting behaviour not influenced by awareness of a camera.[33,37,39]	G >0.8 is achieved with 12 cases assessed[37,39]	Time: 2.5 hours[37] Costs: £268 (€394)[37]
Peer assessment	Content validity: confirmed in 4 studies[41,48,59,63] Construct validity: tested by applying instrument to different populations[45,63] Criterion validity: positive correlations found with knowledge test and faculty evaluations[49,51,58,62]	G> 0.8 is achieved with 7 and 11 raters[43,50]	Time: 1 hour[63] Costs: £107 (€158)[41]
Portfolio or appraisal	Content validity: not yet established, content is considered valid by participants[65] Criterion validity: correlation with annual interview r=0.25[68]	G > 0.8 is achieved with 5 raters[68]	Time: 15-24 hours to compose appraisal folder/portfolio[69,70] Costs: £771 (€1135) per appraisal[70]
Direct observation	Content validity: proven in 2 studies[75,77] Construct validity: demonstrated in 2 studies with doctors with different levels of expertise[75,77]	G= NR Inter-rater reliability r= 0.56[75]	NR
Audit of medical records	Content validity: confirmed for referral letter instrument[81] Not confirmed for medical records: actual performance not properly recorded in 68% of cases, kappa with direct observation 0.12-0.89[78, 83]	G>0.8 with 5 raters and 10 cases (referral letters)[81] or 3 raters and 60 cases (radiologist reports)[79]	Time: 2.5-3.5 hours[70] Costs: NR

G= generalisability coefficient, NR= not reported

VALIDITY

A well performed validity analysis of an instrument should consider different statistical scales, namely, those of: content validity; construct validity, and criterion validity.[91] As can be seen from the detailed summary, hardly any method comprises all aspects of a professional's performance. However, researchers have conducted many studies to establish (part of) the validity of the instruments used. An overview of the most important findings is available in Table 1.

In SP assessments, the content validity is considered indisputable as long as the SP is not detected. Reported detection rates are as low as 1 percent and usually not higher than 8 percent.[27,30] Criterion validity of SPs is not yet established; in 1 study the performance measured by SPs negatively correlated with results on a computerised, case-based test.[29]

Instruments used for video observation have proven content validity.[34–36] Moreover, research has shown that the consulting behaviour of the majority of GPs is not influenced by awareness of a camera in the consulting room and the tapes recorded comprise a representative sample which supports the content validity of video observation.[33,37,39]

Four studies investigated the content validity of peer assessment instruments.[41,48,59,63] Archer et al,[63] Hall et al,[41] Weaver et al,[59] and Van de Camp et al[48] confirmed the content validity of their peer assessment instruments by checking the opinion of experts concerning the composition of the scoring list. Other studies in peer assessment did not report criteria for inclusion of the different items of the instruments. Some evidence for construct validity of a peer assessment instrument is provided by factor analysis, which can show the ability of the instrument to discriminate among experience and specialty differences.[45,63] Positive correlations between peer assessment and faculty evaluations or knowledge tests were found, which gives an indication of the criterion validity of peer assessment.[49,51,62]

The content validity of portfolio or appraisal folders is only supported by the fact that participants consider the content valid because of the focus on personal needs.[65] The construct validity of the portfolio is weakly supported by the correlation of r=0.25 with an annual interview.[68]

Two studies in direct observation determined the content validity and construct validity of an instrument to guide and assess trainee performance in ward rounds.[75,77]

A study in medical audit provided evidence for the content validity and construct validity of an instrument to rate referral letters.[81] Other studies in medical audit do not support the content validity of assessing medical records. Actions undertaken during consultations are often not properly recorded in medical records.[78,83]

Utility of methods as formative assessment

The success of formative assessments is determined by the effectiveness of the method(s) applied. Using a modified version of Kirkpatrick's model, 4 levels of effectiveness can be

observed, with learner satisfaction being the lowest level and improvement in patient and health outcomes the highest.[25] Different levels of effectiveness were the subject of study in 21 of the 58 different studies (Table 2).

Table 2. Effectiveness of performance assessment

Formative assessment (modified version of Kirkpatrick's model)	
1. Learner satisfaction (reaction), 8 studies	
Peer assessment:	-Provides valuable and useful feedback [46,47,50,60]
Portfolio & appraisal:	-83% felt supported[67], majority felt encouraged[70]
Direct observation:	-Worthwhile exercise[76]
2. Learning outcomes (learning), 4 studies	
Peer assessment:	-Leads to formulation of 9 learning objectives[48]
Portfolio & appraisal:	-Encourages continuous professional development [65,69] -Majority reported that portfolio helps in achieving learning objectives[64,65]
3. Performance improvement, *12 studies*	
a) Reported change	
Peer assessment:	-61-72% of doctors reports to initiate a change in behaviour[41,43,47,60,61] -70% of portfolio users indicates to have become more reflective [71,72]
Portfolio & appraisal:	-Reported change in updating medical bags and improved record keeping [69]
b) Measured change	
Medical audit:	-75% of doctors that needs help is successful in improving. [85] -Referral letters improve significantly following feedback[80] -64% of doctors who received a good grade after the first visit, received a lower grade after the second [86]
Portfolio & appraisal:	-No significant increase in portfolio scores from year 1 to 2.[68]
4. Patient/ health outcomes: 0 studies	

QUALITY OF STUDIES EVALUATING EFFECTIVENESS

A quality analysis revealed that the methodological quality of the studies varied considerably. Five studies showed good quality in evaluating the effectiveness of methods.[65,70,85,86,88] Other studies had poor to moderate quality. The poor quality was attributable to several reasons.

Firstly, in 14 of the 21 studies, doctor participation was voluntary. Secondly, changes in routine practice were investigated by self-reporting by doctors.

Thirdly, most studies were conducted in small populations. The number of participating doctors in the studies ranged from 7 to 707 (a total of 3486 doctors in 21 studies). Finally, studies measuring performance improvement lack control groups and effects measured may reflect to the statistical phenomenon of regression to the mean.

EFFECTIVENESS

In 19 studies positive effects were reported, whereas in 2 studies no effect was reported.
Eight studies reported on effectiveness in terms of learner satisfaction (level 1).[46,47,50,60,67,70,76,88] Four studies investigated the effect on learning outcomes (level 2).[48,64,65,69] Achievement of performance improvement was investigated in 12 studies (level 3).[41,43,47,60,61,67,69,71,72,80,85,86] Level 3 can be divided into reported improvements in performance (level 3a) and measured improvements in performance (level 3b). No studies were found concerning effectiveness in terms of patient and health outcomes (level 4).
Performance improvement is usually shown by doctors self-reporting about whether or not they changed their behaviour following the results of the assessment. Doctors involved in peer assessment indicate positive effects for levels 1, 2 and 3. The feedback is valued[46,47,50,60]; leads to the formulation of learning objectives[48] and 61-72 percent of doctors report a change in their behaviour.[43,47,60,61] Research into portfolio and appraisal reports positive effects for levels 1, 2 and 3a. The majority of portfolio users feel encouraged and supported in their professional development[67,70] and report that portfolio helps in achieving learning objectives.[64,65] Reported changes in performance concerned: being more reflective[71,72]; updating medical bags, and improved record keeping.[69] One study demonstrated the absence of any effect in level 3b: there was no significant increase in portfolio scores in the following years.[68] Two studies demonstrated the effectiveness of medical audit. Six years after the first intervention, 75 percent of all underperforming doctors were successful in improving their performance and revisited doctors were practising significantly better.[85] Moreover, referral letters improve significantly following feedback.[80] One study in medical audit demonstrated a negative effect, where 64 percent of assessed doctors showed a decline in grade.[86]

Discussion and conclusion

Relatively few rigorous studies have developed methods of assessing doctor performance. Their science is relatively weak, and very few indeed have combined perspectives - as they should - in order to really measure a construct as complex as doctor performance. Therefore, despite the rhetoric around this area, few real investments have been made. If management is serious about assessing employee performance, then it is high time it got better at doing it. This paper, by way of an extensive systematic review, contributes to the much-needed foundation on methods and instruments for studying and improving doctor performance. Our systematic review of the literature succeeded in identifying a large number of methods and instruments for the performance assessment of individual doctors. The methods and instruments varied greatly in feasibility, reliability, validity and effectiveness. As stated earlier, insights into these items are vital for policy-makers and researchers alike. From a feasibility point of view, we recommend using peer assessment. Reliable results can be achieved with 1 hour of administrative time. This also explains why peer assessment is the form of assessment applied most often in daily practice. Looking at the effectiveness of included methods, we suggest using peer assessment and

portfolio or appraisal in formative assessments, despite the significant shortcomings in quality of the included studies. This is because the majority of doctors subjected to peer assessment and portfolio or appraisal are satisfied with their evaluation and report performance improvements. We consider this an important argument in support of a preference for peer assessment and portfolio or appraisal above other methods.

Policy-makers intending to carry out assessments using SPs or video observation must realise that these approaches are expensive and time-consuming and their effectiveness has not yet been properly studied. None of the methods can be said to be valid from every perspective or for all intents and purposes. To overcome this, policy-makers should aim to incorporate information from multiple sources and various occasions to evaluate the broad spectrum of performance.[93]

We found only 2 performance assessment systems that meet these recommendations. These were a programme that combined portfolio and multisource feedback for junior doctors in the UK and a system comprising audit of medical records, direct observation and portfolio for underperforming doctors in the UK.[88,90] Data from multiple sources are hardly ever combined to evaluate a doctor. This is opposed to recommendations in the literature.

The present review has several limitations. Firstly, it was restricted to English-language publications only, which means that publication bias cannot be ruled out. Secondly, the literature in medical education often lacks the use of extensive medical subject headings, which could have contributed to the non-retrieval of some studies. Finally, the methodological quality of the studies was found to vary greatly and the results should be interpreted with caution. Most of the studies included had been conducted on small, volunteer-based samples.

Future research designs need to pay special attention to unmasking the effectiveness of formative assessments in terms of performance improvement. The concept that assessment drives learning is increasingly acknowledged in medical education as representing a primary principle of good practice in assessment.[94] In our opinion this concept should be extrapolated to clinical care. Empirical evidence supporting improvement in the routine practice of doctors undergoing assessments is lacking. Outcomes of learning of doctors can be determined in terms of observed changes in practice, rather than self-reported changes by doctors.

To ensure that formative performance assessments lead to a change in a doctor's conduct and quality of care, policy-makers should pay more attention to the delivery of feedback and organisational support. Negative and discrepant feedback does not motivate positive change[46] and doctors can experience stress when going through the process of assessment.[95] Research among business managers showed that organisational support is the most important factor for the acceptance of negative 360-degree feedback.[96,97] Thus, hospital organisations should take such issues into account by incorporating mentorship for doctors, organising vocational training for appraisers and mentors of doctors, and responding to issues such as excessive workload or inadequate resources.[98] Given the increasing interest in doctor performance in the literature and in health policy, we expect future methodological and policy advances in this field.

References

1 Grol R. Improving the quality of medical care – building bridges among professional pride, payer profit, and patient satisfaction. JAMA 2001;286:2578–85.

2 Landon BE, Normand SL, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. JAMA 2003;290:1183–9.

3 Peabody FW. The care of the patient. JAMA 1927;88:877–82.

4 Brennan TA, Horwitz RI, Duffy FD, Cassel CK, Goode LD, Lipner RS. The role of physician specialty board certification status in the quality movement. JAMA 2004;292:1038–43.

5 Steinbrook R. Renewing board certification. N Engl J Med 2005;353:1994–7.

6 Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med 2005;142:260–73.

7 Thomson O'Brien MA, Freemantle N, Oxman AD, Wolf F, Davis DA, Herrin J. Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev 2001;2:CD003030.

8 van der Vleuten CPM. The assessment of professional competence: theoretical developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996;1:41–67.

9 Cunnington JPW, Hanna E, Turnbull J, Kaigas TB, Norman GR. Defensible assessment of the competency of the practicing physician. Acad Med 1997;72:9–12.

10 Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. JAMA 1993;270:1046–51.

11 Finucane PM, Bourgeois-Law GA, Ineson SL, Kaigas TM. A comparison of performance assessment programs for medical practitioners in Canada, Australia, New Zealand, and the United Kingdom. Acad Med 2003;78:837–43.

12 Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002;287:226–35.

13 Rethans JJ, Sturmans F, Drop R, van der Vleuten CPM, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. BMJ 1991;303:1377–80.

14 Hays RB, Davies HA, Beard JD, et al. Selecting performance assessment methods for experienced physicians. Med Educ 2002;36:910–7.

15 Scoles PV, Hawkins RE, LaDuca A. Assessment of clinical skills in medical practice. J Contin Educ Health Prof 2003;23:182–90.

16 McKinley RK, Fraser RC, Baker R. Model for directly assessing and improving clinical competence and performance in revalidation of clinicians. BMJ 2001;322:712–5.

17 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. BMJ 2004;328:1240–5.

18 Saturno PJ, Palmer RH, Gascon JJ. Physician attitudes, self-estimated performance and actual compliance with locally peer-defined quality evaluation criteria. Int J Qual Health Care 1999;11:487–96.

19 Reed D, Price EG, Windish DM, et al. Challenges in systematic reviews of educational intervention studies. Ann Intern Med 2005;142:1080–9.

20 Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215–8.

21 Paterson Davenport LA, Hesketh EA, Macpherson SG, Harden RM. Exit learning outcomes for the PRHO year: an evidence base for informed decisions. Med Educ 2004;38:67–80.

22 Tamblyn R, Benaroya S, Snell L, McLeod P, Schnarch B, Abrahamowicz M. The feasibility and value of using patient satisfaction ratings to evaluate internal medicine residents. J Gen Intern Med 1994;9:146–52.

23 Jackson JL, Kroenke K. Patient satisfaction and quality of care. Mil Med 1997;162:273–7.

24 Evans RG, Edwards A, Evans S, Elwyn B, Elwyn G. Assessing the practicing physician using patient surveys: a systematic review of instruments and feedback methods. Fam Pract 2007;29:117–27.

25 Curran VR, Fleet L. A review of evaluation outcomes of web-based continuing medical education. Med Educ 2005;39:561–7.

26 Best Evidence Medical Education Steering Group. Guide for topic review groups on carrying out systematic reviews. http://www.bemecollaboration.org. [Accessed February 2003.]

27 Gorter S, Van der Linden S, Brauer J, et al. Rheumatologists performance in daily practice. Arthritis Rheum 2001;45:16–27.

28 Gorter S, Rethans JJ, Van der Heijde D, et al. Reproducibility of clinical performance assessment in practice using incognito standardised patients. Med Educ 2002;36:827–32.

29 Schuwirth L, Gorter S, Van der Heijde D, et al. The role of a computerised case-based testing procedure in practice performance assessment. Adv Health Sci Educ Theory Pract 2005;10:145–55.

30 Beaulieu MD, Rivard M, Hudon E, Saucier D, Remondin M, Favreau R. Using standardised patients to measure professional performance of physicians. Int J Qual Health Care 2003;15:251–9.

31 Norman GR, Neufeld VR, Walsh A, Woodward CA, McConvey GA. Measuring physicians' performances by using simulated patients. J Med Educ 1985;60:925–34.

32 Cox J, Mulholland H. An instrument for assessment of videotapes of general practitioners' performance. BMJ 1993;306:1043–6.

33 Herzmark G. Reactions of patients to video recording of consultations in general practice. Br Med J (Clin Res Ed) 1985;291:315–7.

34 Tate P, Foulkes J, Neighbour R, Campion P, Field S. Assessing physicians interpersonal skills via videotaped encounters: a new approach for the Royal College of General Practitioners Membership examination. J Health Commun 1999;4:143–52.

35 Enzer I, Robinson J, Pearson M, Barton S, Walley T. A reliability study of an instrument for measuring general practitioner consultation skills: the LIV-MAAS scale. Int J Qual Health Care 2003;15:407–12.

36 Robinson J, Walley T, Pearson M, Taylor D, Barton S. Measuring consultation skills in primary care in England: evaluation and development of content of the MAAS scale. Br J Gen Pract 2002;52:889–93.

37 Ram P, Grol R, Rethans JJ, Schouten B, van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. Med Educ 1999;33:447–54.

38 Hobma SO, Ram PM, Muijtjens AM, Grol RP, van der Vleuten CPM. Setting a standard for performance assessment of doctor)patient communication in general practice. Med Educ 2004;38:1244–52.

39 Hays R, Spike N, Sen GT, Hollins J, Veitch J. A performance assessment module for experienced general practitioners. Med Educ 2002;36:258–60.

40 Campbell LM, Howie JG, Murray TS. Use of videotaped consultations in summative assessment of trainees in general practice. Br J Gen Pract 1995;45:137–41.

41 Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. CMAJ 1999;161:52–7.

42 Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med 1997;72 (Suppl):82–4.

43 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. BMJ 2003;326:546–8.

44 Higgins RSD, Bridges J, Burke JM, O Donnell MA, Cohen NM, Wilkes SB. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. Ann Thorac Surg 2004;77:12–7.

45 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med 2004;79 (Suppl):5–8.

46 Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multi-source feedback: perceptions of credibility and usefulness. Med Educ 2005;39:497–504.

47 Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and co-worker raters to a multi-source feedback process: a pilot study. Acad Med 2003;78 (Suppl):42–4.

48 Van de Camp K, Vernooij-Dassen M, Grol R, Bottema B. Professionalism in general practice: development of an instrument to assess professional behaviour in general practitioner trainees. Med Educ 2006;40:43–50.

49 Notzer N, Eldad A, Donchin Y. Assessment of physician competence in pre-hospital trauma care. Injury 1995;26:471–4.

50 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655–60.

51 Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. Ann Intern Med 1989;110:719–26.

52 Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practising physicians. Acad Med 1996;71:364–70.

53 Brienza RS, Huot S, Holmboe ES. Influence of gender on the evaluation of internal medicine residents. J Womens Health 2004;13:77–83.

54 Rosenbaum ME, Ferguson KJ, Kreiter CD, Johnson CA. Using a peer evaluation system to assess faculty performance and competence. Fam Med 2005;37:429–33.

55 Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynaecology residents. Obstet Gynecol 2002;99:647–51.

56 Elwyn G, Lewis M, Evans R, Hutchings H. Using a 'peer assessment questionnaire' in primary medical care. Br J Gen Pract 2005;55:690–5.

57 Butterfield PS, Mazzaferri EL. A new rating form for use by nurses in assessing residents humanistic behaviour. J Gen Intern Med 1991;6:155–61.

58 Woolliscroft JO, Howell JD, Patel BP, Swanson DB. Resident–patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, programme supervisors, and nurses. Acad Med 1994;69:216–24.

59 Weaver MJ, Ow CL, Walker DJ, Degenhardt EF. A questionnaire for patients evaluations of their physicians' humanistic behaviours. J Gen Intern Med 1993;8:135–9.

60 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med 2002;77 (Suppl):64–6.

61 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. Acad Med 1999;74:702–14.

62 Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. J Gen Intern Med 1999;14:551–4.

63 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. BMJ 2005;330:1251–3.

64 Challis M, Mathers NJ, Howe AC, Field NJ. Portfoliobased learning: continuing medical education for general practitioners – a mid-point evaluation. Med Educ 1997;31:22–6.

65 Mathers NJ, Challis MC, Howe AC, Field NJ. Portfolios in continuing medical education – effective and efficient? Med Educ 1999;33:521–30.

66 McKinstry B, Peacock H, Shaw J. GP experiences of partner and external peer appraisal: a qualitative study. Br J Gen Pract 2005;55:539–43.

67 Boylan O, Bradley T, McKnight A. GP perceptions of appraisal: professional development, performance management, or both? Br J Gen Pract 2005;55:544–5.

68 Melville C, Rees M, Brookfield D, Anderson J. Portfolios for assessment of paediatric specialist registrars. Med Educ 2004;38:1117–25.

69 Bruce D, Phillips K, Reid R, Snadden D, Harden R. Revalidation for general practitioners: randomised comparison of two revalidation models. BMJ 2004;328:687–91.

70 Lewis M, Elwyn G, Wood F. Appraisal of family doctors: an evaluation study. Br J Gen Pract 2003;53:454–60.

71 Campbell C, Parboosingh J, Gondocz T, Babitskaya G, Pham B. A study of the factors that influence physicians' commitments to change their practices using learning diaries. Acad Med 1999;74 (Suppl):34–6.

72 Campbell CM, Parboosingh JT, Gondocz ST, et al. Study of physicians use of a software program to create a portfolio of their self-directed learning. Acad Med 1996;71 (Suppl):49–51.

73 Flood SC. Using qualitative self-evaluation in rating physician performance. Fam Pract Manag 1998;5:22–34.

74 Dornan T, Carroll C, Parboosingh J. An electronic learning portfolio for reflective continuing professional development. Med Educ 2002;36:767–9.

75 Norgaard K, Ringsted C, Dolmans D. Validation of a checklist to assess ward round performance in internal medicine. Med Educ 2004;38:700–7.

76 Dowson C, Hassell A. Competence-based assessment of specialist registrars: evaluation of a new assessment of outpatient consultations. Rheumatology 2006;45:459–64.

77 Filho GRD, Schonhorst L. The development and application of an instrument for assessing resident competence during pre-anaesthesia consultation. Anesth Analg 2004;99:62–9.

78 Stange KC, Zyzanski SJ, Smith TF, et al. How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patients' visits. Med Care 1998;36:851–67.

79 Jolly BC, Ayers B, Macdonald MM, et al. The reproducibility of assessing radiological reporting: studies from the development of the General Medical Council's Performance Procedures. Med Educ 2001;35 (Suppl 1):36–44.
80 Fox AT, Palmer RD, Crossley JG, Sekaran D, Trewavas ES, Davies HA. Improving the quality of outpatient clinic letters using the Sheffield Assessment Instrument for Letters (SAIL). Med Educ 2004;38:852–8.
81 Crossley GM, Howe A, Newble D, Jolly B, Davies HA. Sheffield Assessment Instrument for Letters (SAIL): performance assessment using outpatient letters. Med Educ 2001;35:1115–24.
82 Parkerton PH, Smith DG, Belin TR, Feldbau GA. Physician performance assessment: non-equivalence of primary care measures. Med Care 2003;41:1034–47.
83 Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical performance? A study using simulated patients. Br J Gen Pract 1994;44:153–6.
84 McAuley RG, Henderson HW. Results of the Peer Assessment Program of the College of Physicians and Surgeons of Ontario. CMAJ 1984;131:557–61.
85 Norton PG, Dunn EV, Beckett R, Faulkner D. Longterm follow-up in the Peer Assessment Programme for non-specialist physicians in Ontario, Canada. Jt Comm J Qual Improv 1998;24:334–41.
86 Norton PG, Faulkner D. A longitudinal study of performance of physicians' office practices: data from the Peer Assessment Program in Ontario, Canada. Jt Comm J Qual Improv 1999;25:252–8.
87 Norton PG, Dunn EV, Soberman L. Family practice in Ontario. Can Fam Physician 1994;40:249–56.
88 Hesketh EA, Anderson F, Bagnall GM, et al. Using a 360-degree diagnostic screening tool to provide an evidence trail of junior doctor performance throughout their first postgraduate year. Med Teach 2005;27:219–33.
89 Southgate L, Cox J, David T, et al. The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's Performance Procedures. Med Educ 2001;35 (Suppl. 1):2–8.
90 Southgate L, Cox J, David T, et al. The General Medical Council's Performance Procedures: peer review of performance in the workplace. Med Educ 2001;35 (Suppl. 1):9–19.
91 Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to their Development and Use. Oxford: Oxford University Press, 1989.
92 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. Med Educ 2002;36:972–8.
93 van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ 2005;39:309–17.
94 Handfield-Jones RS, Mann KV, Challis ME, et al. Linking assessment to learning: a new route to quality assurance in medical practice. Med Educ 2002;36:949–58.
95 Cohen DA, Rhydderch M. Measuring a doctor s performance: personality, health and well-being. Occup Med 2006;56:438–41.
96 Brett JF, Atwater LE. 360-degree feedback: accuracy, reactions, and perceptions of usefulness. J Appl Psychol 2001;86:930–42.
97 Facteau CL, Facteau JD, Schoel LC, Russell JEA, Poteet ML. Reactions of leaders to 360-degree feedback from subordinates and peers. Leadersh Quart 1998;9:427–48.
98 Waller DG. Consultant appraisal: pitfalls and how to avoid them. Clin Med 2003;3:569–72.

Chapter 3

Evaluation of doctors' professional performance: An iterative development and validation study of multisource feedback instruments

Karlijn Overeem
Hub C. Wollersheim
Onyebuchi A. Arah
Juliette K. Cruijsberg
Richard P.T.M. Grol
Kiki M.J.M.H. Lombarts

Submitted.

Abstract

Background

In view of demands for high quality care, there is a global need to assess doctors' professional performance in actual clinical practice. Valid and reliable instruments are necessary to support these efforts. This study focuses on the psychometric properties of instruments used for the multisource assessment of doctors' professional performance in the Netherlands.

Methods and findings

This observational validation study of three instruments underlying multisource feedback (MSF) was set in 26 non-academic hospitals in the Netherlands. In total, 146 doctors (internal medicine and surgeons) took part in the study. Their professional performance was assessed by peers (doctor colleagues), coworkers (including nurses, secretary assistants and other healthcare professionals) and patients. Doctors also completed a self-evaluation. Ratings of 864 peers, 894 coworkers and 1960 patients on MSF were available. We used exploratory factor analysis, inter-item correlations, reliability coefficient alpha, inter-scale correlations, and generalisability studies to evaluate the reliability and validity of instruments. Potential biasing factors such as specialty, gender and age were explored with a linear mixed-effects model. We also used Pearsons' correlation coefficient to explore the relation between the three perspectives' and self ratings. Reliability was explored using two methods including G-studies. Factor analysis revealed six, three and one scale with high internal consistency for the peer, coworker and patient questionnaire respectively (Cronbach's alpha 0.95 - 0.96). It appeared that only 2 percent of variance in the mean ratings could be attributed to one influencing factor (member of specialty group). Other factors such as gender of the rater and length of the working relationship did not appear to influence ratings. Self-ratings were not correlated with peer, coworker or patient ratings. However, ratings of peers, coworkers and patients were found to be correlated. Five colleague evaluations, five coworker evaluations and 11 patient evaluations are required to achieve reliable results (reliability-coefficient of 0.70).

Conclusions

The study demonstrates that the three MSF instruments are reliable and valid for evaluating doctors' professional performance in the Netherlands. Scores from peers, coworkers and patients were not correlated with self-evaluations. Future research should examine improvement of performance when using MSF.

Introduction

In view of demands for high quality care, many health care systems aim to assess doctors' professional performance. As the ability to self-assess has shown to be limited[1], there is a need for external assessments. Reliable, valid, feasible and effective measures of performance are vital to support these efforts. Multisource feedback (MSF) or 360-degree evaluation is a relatively new tool which has been studied around the world as a way of assessing multiple components of professional performance. MSF involves external evaluation of performance on various tasks by: 1) peers with knowledge of a similar scope of practice, 2) non-doctor coworkers (nurse, allied healthcare professionals or administrative staff) and 3) patients.[2] Respondents in those three categories who have been able to observe a doctor's behaviour complete questionnaires about a doctor's performance. Doctors themselves also complete a questionnaire about their own performance and these ratings are compared with others' ratings in order to examine needs for change.[3] Before the widespread use of MSF is merited, it is of vital importance that doctors, managers and patients have confidence in the validity and reliability of instruments applied in MSF.[4] In Canada and the United Kingdom, the psychometric properties of questionnaires used for MSF have been studied across different specialties.[5-10] However, Evans et al identified that instruments developed to date lack evidence of validity supporting their use.[11] Furthermore, a recent review on questionnaires designed to gather feedback from patients concluded that few had undergone rigorous reliability and validity testing.[12] In addition, it has been underlined recently that instruments validated in one setting should not be used in new settings without revalidation and updating since validation is an ongoing process, not a one-time event.[13] Hence, given the significance of the judgments made, in terms of both patient safety and the usefulness of MSF for doctors' professional development, it is essential to develop and validate assessment instruments as rigorously as possible. This paper reports on the validation study of three MSF measurement instruments, namely peer completed, coworker-completed and patient-completed. Specifically, this paper addresses three aspects of validity and reliability: (1) the initial psychometric properties of three new instruments based on existing MSF instruments, (2) the relationship between the different instruments including self-evaluation, (3) the number of evaluations needed per doctor to establish the reliability of assessments.

Methods

MSF-system in the Netherlands

The MSF system in the Netherlands consists of feedback from doctor colleagues (peers), coworkers and patients. This is combined with a reflective portfolio and an interview with a trained mentor (a colleague from a different specialty based in the same hospital) to increase the acceptance of feedback and the chance of performance improvement. To guide future performance, the mentor helps doctors interpret the feedback and critically analyze their performance making use of the feedback. As part of a larger doctors' performance project the MSF-system was launched for the assessment of medical specialists' performance in 2007 in

three hospitals and a pilot study established its feasibility.[14] Subsequently, the MSF system has been adopted by 23 other hospitals. Since 2010, participation in doctor performance assessments is a performance indicator for the Healthcare Inspectorate. The MSF process is managed electronically by an independent web service. Specialists are invited via e-mail and asked to complete a self-evaluation form and nominate up to 16 raters (8 peers and 8 coworkers). All raters except patients are contacted by e-mail and are asked to complete a questionnaire via a dedicated web portal protected by a password login. Data collection from patients takes place via paper questionnaires. The web service automatically sends reminders to non-respondents after 2 weeks. Consecutive patients attending the outpatient clinic of the doctor participating are offered the questionnaire by the receptionist on arrival. Patients are asked to complete the questionnaire after the consultation and to post it in a sealed box. The web-based service provides electronic feedback reports to the mentor and doctor to be discussed face-to-face in a personal interview. The report contains global overall graphic and detailed numeric outcomes of the peers, coworkers and patients' evaluations. Free text comments (answers from raters to open questions) are also provided at the end of the MSF report.

MSF instrument and development

There were two distinct stages of instrument development as part of the validation study. The two stages are described below.

CONTENT GENERATION AND CONTENT VALIDITY. The research committee (5 members) drafted a questionnaire and drew on previously developed MSF instruments in Canada.[2] The 20 items of the patient questionnaire that concerned management of the practice (such as performance of staff at the outpatient clinic) were removed as the aim of the project was to measure doctors' professional performance and those items are the subject of another system.[15] Two researchers translated the items of the questionnaires. A backward translation-check was performed by an independent third person. Next, content validity was established in a small study. Fifteen doctors, ten coworkers and ten patients were asked to rate the relevance and clarity of questions on a 1 to 4 scale. (1=not relevant/not clear, 4=very relevant/very clear). The accepted norm for inclusion of an item in its current format was if 70 percent of respondents agreed (a score of 3 or 4). For the peers' and coworkers' questionnaires, all original items were found to be relevant; 6 items on the peer questionnaire needed reformulation for the purpose of clarification. Two items were removed from the patient questionnaires as they were perceived as irrelevant for the Dutch context and eight items of the patient questionnaire needed reformulation for clarity.

PILOT FIELD TESTING. In total, 45 doctors participated in a pilot study to investigate the feasibility of the system and appropriateness of items. The feasibility results are described elsewhere.[14] The *appropriateness* of items was evaluated through the item-response frequencies. An item was

judged suitable for the MSF questionnaire if at least 60 percent of the raters (peers, coworkers or patients) responded to the item. After analysis of items with a >40 percent category of 'unable to evaluate', five items were removed from the peer questionnaire and two items were removed from the patient questionnaire.

FINAL MSF SYSTEM. The final MSF system used in the study presented in this paper comprised three questionnaires, each prefaced by an introduction. The peer questionnaire consisted of 33 performance items; the coworker and patient questionnaires included 22 and 18 items respectively. All items invited responses on a 9-point Likert type scale: (1=completely disagree, 5=neutral, 9=completely agree). On every item, raters had the option to fill in: 'unable to evaluate'. In addition, all raters were asked to fill in two open questions for narrative feedback, listing the strengths of individual doctors and formulating concrete suggestions for improvement.

Study design, population and setting

This observational validation study on the use of three MSF instruments in actual practice was set in 26 non-academic hospitals in the Netherlands, including both surgical and non surgical specialties. For several specialties such as anaesthesiology and radiology different instruments were developed[5,16] and therefore excluded from our study. All doctors who completed the interview with a mentor were approached to participate. No financial incentives were provided and participants could withdraw from the study at any time without any consequences. Participating doctors consented to provide their anonymous data for research analysis. We aimed to obtain a large sample with sufficient data (more than 100 respondents) to allow an assessment of the performance of the questionnaires in line with recognised best practice.[13] Data collection took place in the period September 2008-July 2010. The analysis presented in this paper used anonymised datasets derived from this volunteer sample. The study was approved by the Institutional Review Board as an expedited approval since the participants in our study are not patients.

Statistical analysis

We conducted exploratory factor analysis, reliability coefficients, item-total scale correlation and interscale correlations. For item reduction we conducted exploratory factor analyses (extraction method: principal components technique; extraction criterion: eigenvalue > 1; rotation method: varimax rotation) to explore the factor or scale structure underlying the questionnaires. Items were assigned to the scale on which they loaded with at least a factor loading of 0.30 (to avoid low-loading items and in line with the literature). In the case of cross factor loadings, an item was assigned to where it loaded the highest factor unless it was theoretically appropriate to leave it under the factor on with it loaded the second highest. Next, each composite-scale was calculated as an average of the items that loaded the highest on it. Subsequently, the scale structure was subjected to reliability analysis using Cronbach's alpha. We considered a Cronbach's alpha of at

least 0.70 as an indication of satisfactory internal consistency reliability of each scale. To investigate reliability further, we checked for homogeneity of scales by examining the item-total correlations corrected for overlap. Item-total scale correlations of 0.40 or higher were considered acceptable evidence of contribution of each item to the scale homogeneity. We further assessed the degree of overlap between scales by estimating inter-scale correlations using Pearsons' correlation coefficient. Like elsewhere, correlations of <0.7 between the scales was considered appropriate.[13] Second, to quantify the potential influences on the doctors' ratings, we built a model that controlled for the effect of the individual doctor, and the bias with which an individual rater (peer, coworker or patient) rates the doctor. To accomplish this, we used a linear mixed-effects model to look at the adjusted estimate of each factor while correcting for the nesting of raters within specialists. As independent variables, we included gender of the rater, length of the professional relationship of rater, specialty, work experience of the doctor, gender of the doctor and membership of the same specialist group as independent variables. Subsequently, we examined the relationships between the four measurement perspectives (self, peer, coworker and patient) using Pearsons' correlation coefficient. For the estimation of these relationships, we used the mean score of all items per rater. Finally, we estimated the number of colleagues, patients and coworkers needed for achieving reliable ratings per doctor. A variety of techniques have been developed for estimating the reliability of ratings. We applied a formula recently described in another journal as well as a generalisability study.[17] Historically, generalisability is used to estimate the number of raters necessary to produce reliable results. Generalisability studies incorporate various sources of error such as variance attributable to the doctor, to the rater and error variance. In our dataset, raters were unique to the doctors. This fully nested model allows only the estimation of two variance components: true variance (attributable to the doctor) and residual variance (all other variance). Generalisability is then calculated as follows:

True variance/ [True variance + (error variance/n)]

Results

Study participants

A total of 146 specialists participated in the study. In total 864 peers (a mean of 6.5 per doctor) 894 coworkers (a mean of 6.7 per doctor) and 1890 patients (a mean of 15 per doctor) rated the specialists. Forty percent of the doctor participants was female. The mean number of years since first registration of the doctors was 13.6 years, (minimum 2 years; maximum 35 years; standard deviation 8.4 years). Of the raters, 35 percent of peers, 81 percent of coworkers and 65 percent of the patients were female.

Mean ratings and missing data

Peers scored doctors highest on the items 'responsibility for patients' (mean= 8.67) and 'responsibility for own professional actions' (mean= 8.64). Peers provided the lowest ratings for

the item 'research activities' (mean= 7.67) and 'evaluating literature' (mean= 7.96). When aggregated for the individual doctor, the mean rating given by peers was 8.37, ranging from 7.67 (min 1 max 9 SD 1.75) to 8.69 (min 2 max 9 SD 0.70). All items were positively skewed. Coworkers rated doctors highest on 'responsibility for professional actions' (mean= 8.64) and lowest on 'verbal communication with coworkers' (mean= 7.78). Patients rated doctors highest on 'respect' (mean= 8.54) and gave doctors the lowest rating for 'asking details about personal history' (mean= 7.72). Missing data (unable to comment) ranged from 4 percent of coworkers' responding on the item 'collaborates with doctor colleagues' to 38.9 percent of peers evaluating doctors' performance on 'participates adequately in 'research activities'. On average, per item, the mean of missing data was 19.3 percent for peers, 10 percent for coworkers' responses and 17.7 percent for patients. All mean scores of items are summarised in table 1A, B and C.

Table 1A. Factors derived from the principal components analysis of colleagues' ratings

Scale and items	Mean score [SD]	Factor loadings on primary scale	Internal consistency reliability	Corrected item-total correlations
Collaboration and self-insight [42% of variance]	8.47 [1.09]		0.900	
Communicates effectively with other health care professionals	8.23 [1.07]	0.581		0.655
Collaborates with doctor colleagues	8.56 [0.91]	0.841		0.795
Accepts feedback provided	8.38 [1.06]	0.748		0.711
Recognises his/her own limitations	8.13 [1.19]	0.643		0.702
Participates effectively as a member of the health care team	8.40 [1.21]	0.631		0.700
Exhibits professional behaviour towards doctor colleagues	8.61 [0.85]	0.779		0.750
If a member of my own family needed care I would recommend this doctor	8.59 [0.84]	0.760		0.761
Clinical performance [8% of variance]	8.40 [0.79]		0.900	
Performs technical procedures skilfully	8.45 [0.96]	0.638		0.615
Selects diagnostic tests appropriately	8.38 [0.93]	0.739		0.758
Critically assesses diagnostic information	8.44 [0.93]	0.763		0.823
Makes the correct diagnosis following consultation	8.43 [0.88]	0.780		0.822
Selects appropriate treatments	8.41 [0.93]	0.650		0.791
Accepts responsibility for own professional actions	8.40 [0.91]	0.452		0.629
Coordination and continuity [5% of variance]	8.47 [0.73]		0.851	
Handles transfer of care appropriately	8.37 [1.03]	0.727		0.697
Maintains confidentiality of patients and their families	8.57 [0.89]	0.660		0.601
Provides a clear understanding about who is responsible for the continuing care of patients	8.35 [0.94]	0.632		0.717
Co-ordinates care effectively for patients with other health care professionals and doctors	8.44 [0.89]	0.609		0.684
Maintains quality medical records	8.13 [1.23]	0.456		0.547
Manages patients with complex problems	8.46 [0.88]	0.632		0.669

Practice based learning and improvement [4% of variance]	8.12 [1.13]		0.813	
Contributes to quality improvement programs and practice guidelines	8.22 [1.26]	0.652		0.726
Teaches adequately medical colleagues and coworkers	7.97 [1.42]	0.652		0.728
Participates adequately in research activities	7.67 [1.52]	0.599		0.739
Critically evaluates the medical literature	7.96 [1.36]	0.655		0.725
Emergency medicine [4% of variance]	8.38 [0.85]		0.767	
Gives priority to urgent requests	8.46 [0.93]	0.660		0.634
Handles emergency situations effectively	8.49 [0.91]	0.703		0.631
Manages own stress effectively	8.22 [1.12]	0.564		0.549
Time-management and responsibility [4% of variance]	8.69 [1.30]		0.770	
Handles requests for consultation in a timely manner	8.40 [1.00]	0.749		0.645
Advises referring doctor if referral request is outside the scope of his/her practice	8.53 [0.87]	0.550		0.576
Assumes appropriate responsibility for patients	8.67 [0.69]	0.527		0.539
Provides timely information to referring doctors about mutual patients	8.28 [0.13]	0.690		0.608

Table 1B. Factors derived from the principal components analysis of coworkers' ratings

Scale and items	Mean score [SD]	Factor loadings on primary scale	Internal consistency reliability	Corrected item-total correlations
Relationship with other health care professionals [57% of variance]	8.07 [1.11]		.925	
Is able to verbally communicate effectively with other health care professionals	8.10 [1.21]	0.691		0.757
Is courteous to coworkers	8.35 [1.11]	0.760		0.763
Respects the professional knowledge and skills of coworkers	8.31 [1.10]	0.782		0.765
Collaborates well with coworkers	8.31 [1.10]	0.811		0.848
Is accessible for appropriate communication about patients	8.37 [1.06]	0.611		0.728
Participates effectively as a member of the health care team	8.28 [1.13]	0.660		0.766
This doctor presents him/herself in a professional manner	8.59 [0.92]	0.574		0.740
Communication with patients [7% of variance]	8.03 [1.07]		.900	
Communicates effectively with patients	8.21 [1.15]	0.830		0.794
Communicates effectively with families	8.11 [1.24]	0.818		0.764
Shows compassion to patients and their families	8.36 [1.07]	0.721		0.812
Is courteous to patients and their families	8.56 [0.89]	0.656		0.772
Respects the rights of patients to make informed decisions	8.46 [0.96]	0.578		0.706
Is reasonably accessible to patients	8.28 [1.08]	0.579		0.700
Patient care [6% of variance]	8.29 [1.06]		.830	
Accepts responsibility for patient care	8.64 [0.72]	0.748		0.720
Maintains confidentiality of patients	8.69 [0.77]	0.711		0.613
Accepts responsibility for professional actions	8.64 [0.86]	0.781		0.773
Responds appropriately in emergency situations	8.40 [1.11]	0.643		0.586

Table 1C. Factors derived from the principal components analysis of patients' ratings

Scale and items	Mean score [SD]	Factor loadings on primary scale	Internal consistency reliability	Corrected item-total correlations
Patient-centeredness [60% of variance]			.959	
Explained my illness or concern to me clearly	8.30 [1.32]	0.792		0.792
Spends enough time with me	8.29 [1.38]	0.825		0.781
Shows interest in my problems	8.25 [1.43]	0.840		0.806
Answers my questions well	8.33 [1.32]	0.873		0.827
Treats me with respect	8.54 [1.02]	0.810		0.763
Shows compassion	8.10 [1.53]	0.804		0.770
I would go back to this doctor	8.50 [1.23]	0.819		0.803
I would recommend this doctor to others	8.43 [1.34]	0.828		0.823
Explains my treatment choices or options	8.11 [1.48]	0.771		0.769
Tells me how and when to take my medicine	8.00 [1.59]	0.646		0.722
Explains clearly different steps of my treatment plan (including risks and benefits)	8.06 [1.53]	0.763		0.779
Asks details about my personal history when appropriate	7.72 [1.83]	0.664		0.659
Explains my physical exam clearly	8.13 [1.49]	0.799		0.750
Asks permission for some treatments or exams	8.03 [1.60]	0.689		0.697
Explains clearly what could be done in unsuspected circumstances, such as fever, illness or changes in my complaints	7.84 [1.70]	0.757		0.754
Tells me what to do if my problems do not get better	7.90 [1.23]	0.785		0.787
Makes sure that my other caregivers are well informed	7.94 [1.68]	0.670		0.650

Dimension structure and reliability of the dimensions

Factor loadings from principal components analysis of the peer ratings, yielded 6 factors with an Eigen value greater than 1, in total explaining 67 percent of variance. The factors comprised: collaboration and self-insight, clinical performance, coordination & continuity, practice based learning and improvement, emergency medicine, time management & responsibility. Due to low factor loadings, three items were eliminated. These factors were highly consistent with the structure of the questionnaire, as defined by items having a factor loading greater than 0.4 (Table 2). Principal components analysis of the coworker instrument revealed a 3-factor structure explaining 70 percent of variance. Because of low factor loadings and high frequency of 'unable to evaluate', five items were removed from the instrument. Scales included: relationship with other healthcare professionals, communication with patients and patient care. The principal components analysis of the patient ratings yielded a 1-factor structure explaining 60 percent of the total variance. Cronbach's alphas were high for peers', coworkers' and patients' composite scales, ranging from 0.77 to 0.95. (Table 1A,B,C) Cronbach's alpha for the peer, coworker and patient questionnaires were 0.95, 0.95 and 0.94 respectively, indicating good internal consistency and reliability of the questionnaires. Item-total correlations yielded homogeneity within composite scales. Inter-scale correlations were positive and <0.7, indicating that all the scales of the three instruments were distinct. (see Table 2)

Table 2A. Pearson correlation coefficient between colleagues' scales

	Collaboration	Clinical performance	Practice based learning and improvement	Coordination and continuity	Responsibility and time-management	Emergency medicine
Collaboration	1.000	0.499*	0.451*	0.295*	0.459*	0.437*
Clinical performance		1.000	0.551	0.432*	0.383*	0.408*
Practice based learning and improvement			1.000	0.357*	0.445*	0.400*
Coordination and continuity				1.000	0.338*	0.343*
Responsibility and time-management					1.000	0.328*
Emergency medicine						1.000

* correlation is significant at 0.01 level

Table 2B. Pearson correlation coefficient between coworkers' scales

	Relationship with healthcare professionals	Communications with patients	Patient care
Relationship with healthcare professionals	1.000	0.667*	0.537*
Communications with patients		1.000	0.574*
Patient care			1.000

* correlation is significant at 0.01 level

Factors influencing rating

The linear mixed model showed that membership of the same specialist group is positively correlated with the overall rating given to colleagues. (Beta = 0,153, $p<0.01$). There was a small but significant influence of specialists' work experience, showing that specialists with more experience tend to be rated lower by peers (Beta = -0,008, $p<0.05$) and coworkers (Beta = -0.012, $p<0.05$). These two biasing factors accounted for 2 percent of variance in ratings. Across coworker assessors there was a significant difference in scores on the basis of gender, showing that male coworkers tend to score specialists lower compared to female coworkers.

(Beta = -0.200, $p<0.001$). This factor explained 2 percent of variance. We found no statistical effect of the length of the relationship of the coworkers and peers with the specialist. The patients' age was positively correlated with the ratings provided to the specialist (Beta = 0.005, $p<0.001$). Finally, we found no statistical influence of patients' gender. The model for patient ratings accounted for only 3 percent of the variance in ratings. All parameter estimates of biasing factors are summarised in Table 3.

Table 3. Effects of raters' characteristics and doctors' characteristics on overall mean scores

	Overall rating peers		Overall rating coworkers		Overall rating patients	
	Parameter estimated coefficient [SE]	P-value	Parameter estimated coefficient [SE]	P-value	Parameter estimated coefficient [SE]	P-value
Doctors' characteristics						
Male (reference: female)	-.0137 [.065]	.832	-.019 [.091]	.838	-.049 [.091]	.591
Years of experience	-.008 [.004]	.043*	-.012 [.005]	.029*	.003 [.005]	.598
Surgery	ref.		ref.		ref.	
Internal medicine	.069 [.064]	.287	.139 [.094]	.140	.096 [.088]	.280
Raters' characteristics						
Female	ref.		ref.		ref.	
Male	-.002 [.051]	.974	-.200 [.071]	.005*	-.055 [.062]	.378
Age					.005 [.002]	.002*
Relationship						
Membership of the same specialist group	.153 [.049]	.002*				
Working together: < 6 months	-.115 [.179]	.523	.046 [.185]	.804		
Working together:> 6 months and < 1 year	.042 [.111]	.702	-.036 [.122]	.770		
Working together >1 year	ref.	ref.				

*p< 0.05

Relationship between the different ratings

Self-ratings were not significantly correlated with the peer ratings, coworker ratings or patient ratings. All other mean ratings in the MSF procedure appeared to be correlated. Peer ratings were correlated with the patient ratings (r=0.214, p<0.01). The correlation between the peer ratings and the coworker ratings was significant as well (r=0.352, p<0.01). Finally, coworker ratings appeared to be positively associated with patient ratings. (r=0.220, p<0.01) Table 4 shows the correlations between the mean scores for self ratings, colleague ratings, coworker ratings and patient ratings.

Table 4. Pearsons' correlation coefficients between the ratings of four measurements perspectives: self, colleagues, coworkers and patients

	Self rating	Medical colleagues' ratings	Coworkers' ratings	Patient ratings'
Self rating	1.000	0.062	0.082	0.067
Medical colleagues' ratings		1.000	0.352*	0.214*
Coworkers' ratings			1.000	0.220*
Patient ratings'				1.000

* correlation is significant at 0.05 level

Determining the minimum sample size required

The reliability analysis using the *formula explained* in a previous article[17] showed that - assuming a reliability coefficient of 0.60- ratings from 4 peers, 4 coworkers and 9 patients are required for reliable feedback to doctors. When we would apply a stricter reliability coefficient of 0.70, as many as 5 peers, 5 coworkers and 11 patients evaluating each doctor are required. Analyses using traditional generalisability studies (G-study) revealed that 17 peers, 7 coworkers and 26 patients are necessary to achieve reliable results. The various variance components for this calculation are provided in Table 5.

Table 5. Variance components for the three different groups of raters

	True variance	Residual variance
Peers	0.06	0.42
Coworkers	0.16	0.45
Patients	0.093	1.07

Table 6 summarises the number of raters needed for reliable results based on the two procedures applied.

Table 6. Number of colleagues, coworkers and patients' evaluations needed per doctor for reliable evaluation of doctors' professional performance for different reliability coefficients

	Reliability coefficient of 0.60	Reliability coefficient of 0.70	Reliability coefficient of 0.80
Peers	4, 11	5, 16	5, 28
Coworkers	4, 4	5, 7	6, 11
Patients	9, 17	11, 26	12, 46

Numbers refer respectively to:
(i) extrapolation based on the formula explained in a previous article[17]
(ii) the number of evaluations from peers, coworkers and patients needed per doctor based on generalisability studies with variance components

Discussion

Main findings

This study shows that the adapted Canadian MSF tool, incorporating peer, coworker and patient feedback questionnaires is reliable and valid for hospital based doctors (surgical and non-surgical). Principal components analysis demonstrated good internal consistency and a robust scale structure of the three instruments. We found that little of the variance in performance could be explained by factors outside the doctors' control such as gender of the rater and length of the relationship with the rater. Specialists were rated more positively by members of their specialist group but this accounted for only two percent of variance in ratings. Individual reliable feedback reports can be generated with a minimum of 5 evaluations of colleagues, 5 coworkers and 11 patients respectively, supporting the feasibility of the instruments in Dutch hospitals' settings.

Explanation and interpretation

Our findings provide strong empirical support for the reliability and validity of the results obtained from the three MSF instruments for specialists' performance evaluation. The results of the psychometric analyses for the three MSF instruments indicate that we could tap into multiple scales per questionnaire. For the peer instrument, our factor analysis suggested a 6-dimensional scale. These findings do not support the 4-dimensional structure found in earlier research of the original instruments by Violato and Lockyer. Other studies of instruments used for MSF by Archer et al (2005)[18] and Ramsey et al (1993)[19] assess two generic factors labelled as clinical and psychosocial qualities. Our findings do not confirm the suggestions made in the abovementioned studies. Other researchers argue that in MSF evaluations, the halo effect -which is the tendency to give global impressions- and stereotyping exist.[20,21] This does not seem to apply to Dutch hospital doctors evaluating colleagues. Doctors seem to be able to distinguish between different aspects of professional performance instead of giving global impressions concerning the clinical and humanistic qualities. Our finding that self-ratings using MSF are not related with ratings made by peers, coworkers and patients is consistent with the current literature on self-assessment.[1] We found support for significant correlations between ratings of peers, coworkers and patients. However, correlations were found to be weak. They can be considered as three independent groups of raters, representing different perspectives. Debatably, the fact that their ratings are correlated is a measure of concurrent validity. Similarly with other MSF instruments, we have not formally tested the criterion validity of instruments because a gold standard test of doctors' performance is lacking.[11] Based on generalisability studies, our results suggest that evaluations of 17 colleagues, 7 coworkers and 26 patients per doctor will be needed for generating reliable feedback reports. However, we believe the results of the residual variance are overestimated in nested datasets because no information can be gathered about the variance attributable to the rater. Raters 'grouped in nests' (each nest unique to each doctor) are likely in themselves to vary in their stringency independently of any true performance difference between doctor.[20] Therefore, we believe the results of the formula based on Cronbach's alpha produces more adequate results. The number of 5 peers, 5 coworkers and 11 patients seems attainable for most Dutch specialties and hospitals. As an alternative method, some authors prefer to present reliability as a measure of precision and spread of scores.[22] They calculate 95% CIs by multiplying the SEM (Standard error of measurement) by 1.96 and adding and subtracting this from a mean rating.[20] The CI, generated for the number of raters that contributed to the individual doctor mean score, can then be placed around that score. This provides a measure of precision and, therefore, the reliability that can be attributed to each mean score based on the number of individual scores contributing to it. Given the high scores of doctors, calculations based on 95% CIs displayed that indeed with 5 peers, 5 coworkers and 11 patients none of the doctors scored less than the criterion standard, in our case 6.0 on a 9-point standard.

Strengths and limitations

This study was restricted to a self-selected sample of doctors receiving feedback. It is likely that those who agreed to participate were reasonably confident about their own standards of practice and the sample may have been skewed towards good performance. The mean scores however are similar to scores reported by other but similar instruments which were also skewed to good performance.[19] Nevertheless, we do not expect the statistical findings to be different on validity and reliability with a sample including non-volunteers as these doctors have a similar mix of patients and colleagues. Second, we could use only 80 percent of peer responses due to missing values on one or more items. Future work should investigate whether missing values are indicative of the tendency to avoid a negative judgment. Third, administrative assistants were asked to distribute the survey to consecutive patients at the outpatient clinic but we were not able to check if this was correctly executed for all participants. Finally, because of the cross-sectional design of this study, an assessment of intra-rater (intra-colleague or intra-coworker) or test-retest reliability was not possible. Further work on the temporal stability of responses of the questionnaires is warranted.

Implications for practice and research

This study established the validity and reliability of MSF for medical specialists in the Netherlands. Although it cannot be expected that one single tool can guide improvement for all doctors, it offers doctors in the Netherlands feedback about their performance. MSF in the Netherlands has been designed and tested for formative purposes. The purpose is to give feedback to doctors so that they can steer their personal development plans towards achieving performance excellence. Both the fact that results are reliable with 5 peer, 5 coworkers and 11 patient raters and the shortening of the questionnaire while remaining valid, promises a feasible process. This contributes to the feasibility of the system. Because in the Dutch system also a collegial, time-investing interview takes place, a feasible MSF procedure is of great importance. We did not test the possibility to use the results of our study to draw conclusions about the possibility to detect doctors whose performance might be below standard. In view of the positive skewness of results and the fact that criterion validity is not yet tested, we consider this as an undesirable development. We consider this study as a starting point for further research. We agree with Archer et al that MSF is unlikely to be successful without robust regular quality assurance to establish and maintain validity including reliability.[20] As a result we do not claim the items presented in the tables to be the final version, since a validation process should be ongoing. Furthermore, additional work is required to further establish the validity of the instruments. Further validity of the instruments could be tested by comparing scores with observational studies of actual performance of doctors while conducting their jobs, requiring for example external teams of observers or simulated patients. Finally, it would be useful to determine whether good performers receive significantly less tips for improvement in narrative comments compared to underperforming doctors.

References

1 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of doctor self-assessment compared with observed measures of competence - A systematic review. JAMA 2006;296:1094-102.

2 Hall W, Violato C, Lewkonia R, et al. Assessment of doctor performance in Alberta: the doctor achievement review. CMAJ 1999;161:52-7.

3 Atwater LE, Brett JF. Antecedents and consequences of reactions to developmental 360 degrees feedback. J Vocat Behav 2005;66:532-48.

4 Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. Qual Saf Health Care 2008;17:187-93.

5 Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. Can J Anaesth 2006;53:33-9.

6 Lockyer JM, Violato C, Fidler H. The assessment of emergency doctors by a regulatory authority. Acad Emerg Med 2006;13:1296-303.

7 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. BMJ 2003;326:546-8.

8 Violato C, Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. Pediatrics 2006;117:796-802.

9 Davies H, Archer J, Bateman A, et al. Specialty-specific multi-source feedback: assuring validity, informing training. Med Educ 2008;42:1014-20.

10 Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family doctors and their colleague and coworker raters to a multi-source feedback process: a pilot study. Acad Med 2003;78:S42-S44.

11 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of doctors. BMJ 2004;328:1240-5.

12 Department of Health. Trust, Assurance and Safety: The Regulation of Health Professionals. [British policy document],2007.www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_085162. Accessed on 25-3-2011.

13 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th Edition. Oxford; Oxford university press, 2008: 5-36, 167-206 and 247-74.

14 Overeem K, Lombarts MJ, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. Med Teach 2010;32:141-7.

15 Lombarts MJMH, Klazinga NS. A policy analysis of the introduction and dissemination of external peer review (visitatie) as a means of professional self-regulation amongst medical specialists in The Netherlands in the period 1985-2000. Health Policy 2001;58:191-213.

16 Lockyer JM, Violato C, Fidler HM. Assessment of radiology doctors by a regulatory authority. Radiology. 2008;247:771-8.

17 Lombarts KM, Bucx MJ, Arah OA. Development of a system for the evaluation of the teaching qualities of anesthesiology faculty. Anesthesiology 2009;111:709-16.

18 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. BMJ 2005;330:1251-3.

19 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate doctor performance. JAMA 1993;269:1655-60.

20 Archer J, McGraw M, Davies H. Republished paper: Assuring validity of multisource feedback in a national programme. Postgrad Med J 2010;86:526-31.

21 Borman WC. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. J Appl Psychol 1975;60:556-60.

22 Archer J, Norcini J, Southgate L, Heard S, Davies H. mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? Adv Health Sci Educ Theory Pract 2008;13:181-92.

Chapter 4

Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives

Karlijn Overeem
Kiki M.J.M.H. Lombarts
Onyebuchi A. Arah
Niek S. Klazinga
Richard P.T.M. Grol
Hub C. Wollersheim

Medical Teacher 2010;32:141–147.

Abstract

Background

Doctor performance assessments based on multi-source feedback (MSF) are increasingly central in professional self-regulation. Research has shown that simple MSF is often unproductive. It has been suggested that MSF should be delivered by a facilitator and combined with a portfolio.

Aims

To compare three methods of MSF for consultants in the Netherlands and evaluate the feasibility, topics addressed and perceived impact upon clinical practice.

Method

In 2007, 38 facilitators and 109 consultants participated in the study. The performance assessment system was composed of (i) one of the three MSF methods, namely, Violato's Physician Achievement Review (PAR), the method developed by Ramsey et al for the American Board of Internal Medicine (ABIM), or the Dutch Appraisal and Assessment Instrument (AAI), (ii) portfolio, (iii) assessment interview with a facilitator and (iv) personal development plan. The evaluation consisted of a postal survey for facilitators and consultants. Generalized estimating equations were used to assess the association between MSF method used and perceived impact.

Results

It takes on average 8 hours to conduct one assessment. The CanMEDS roles 'collaborator', 'communicator' and 'manager' were discussed in, respectively, 79, 74 and 71 percent of the assessment interviews. The 'health advocate role' was the subject of conversation in 35 percent of the interviews. Consultants are more satisfied with feedback that contains narrative comments. The perceived impact of MSF that includes coworkers' perspectives significantly exceeds the perceived impact of methods not including this perspective.

Conclusions

Performance assessments based on MSF combined with a portfolio and a facilitator-led interview seem to be feasible in hospital settings. The perceived impact of MSF increases when it contains coworkers' perspectives.

Introduction

Today's doctors are confronted with an explosion of medical knowledge and the need to apply evidence in day-to-day clinical practice. In addition, they have to collaborate in multi-disciplinary, ever larger teams and communicate with well-informed patients. To reflect these changes in practice, doctors need to update their competences continuously to perform optimally.[1] As a result, it seems necessary that doctors are assessed in daily practice to inform them about their performance.

Literature on performance assessment shows that the incorporation of information from multiple sources and various occasions is essential to evaluate a complex construct as doctor performance.[2,3] One of the methods often used in these assessments is multi-source feedback (MSF).[4,5] Research shows that simple feedback is often unproductive.[6,7] It has been recommended to implement a portfolio that stimulates reflection and a facilitator who delivers MSF in order to increase the acceptance of feedback.[7-9] However, it is not yet clear whether these targeted multi-source assessments are feasible in clinical practice and which elements of MSF are critical to an improvement in doctor performance.

In this study, we designed and evaluated a performance assessment system for consultants in the Netherlands. According to the findings from a recent systematic review, we developed a performance assessment system based on MSF.[10] It extends earlier work by incorporating a reflective portfolio and an interview with a collegial facilitator. As it was not yet clear what elements of MSF are critical to the impact of the assessments, we tested three methods of MSF, namely: Violato's Physician Achievement Review (PAR)[11], the method owned by the American Board of Internal Medicine (ABIM), developed by Ramsey et al[12] and the Dutch Appraisal and Assessment Instrument (AAI).[13] In particular we addressed the following research questions:

1. 'What is the feasibility of the performance assessments in hospital settings?'
2. 'What topics are addressed most often in the performance assessment interviews?'
3. 'What are consultants' perceptions of the impact on future clinical practice of the performance assessments and what are the differences between the three MSF methods tested?'

Method

Context

In January 2006, the Dutch Organisation of Medical Specialists – the umbrella organisation of consultants – launched a performance assessment project for consultants in the Netherlands. The project was aimed at developing, testing and evaluating a performance assessment system and at improving doctors' performance. Eight hospitals voluntarily participated in the project. Every hospital appointed a hospital project leader who was responsible for the implementation, organisation and progress of the assessments. Two 1 day meetings with participating consultants and facilitators were held to organise and encourage feedback on methods and instruments and to establish commitment. Hospitals developed their own introduction plans, including the

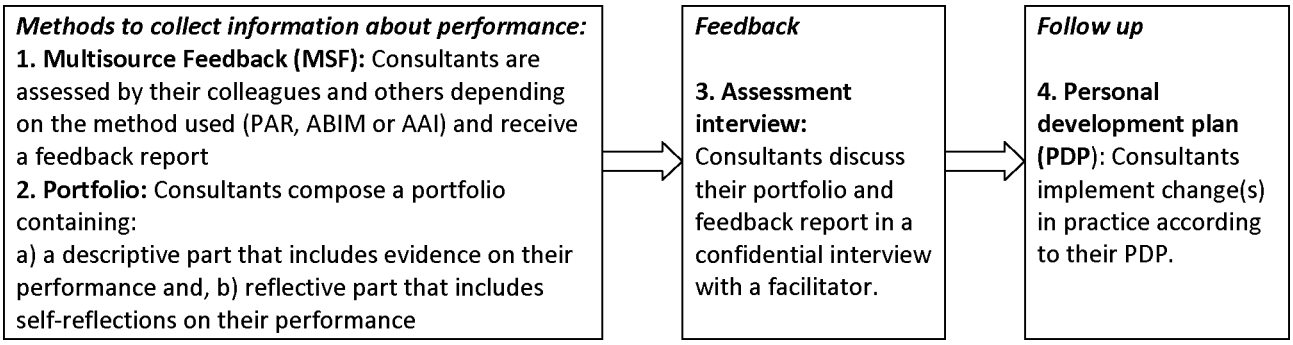
recruitment of collegial facilitators and consultants for voluntary participation in the performance assessment project. Facilitators were offered 1 day of training. The training included explanation of the performance assessment system and procedures and role-plays. Each hospital committed to complete 15 assessments. The assessments were conducted between November 2006 and June 2007. Hospitals selected one MSF method (PAR, ABIM or AAI) based on their preferences and possibilities.

Intervention

The three performance assessment systems compared in this study included the same four components (see Figure 1 for a detailed overview of the assessment system):

- MSF. Every consultant receives feedback from his/her colleagues, coworkers and/or patients using one of the three methods of MSF. The feedback is summarised in a feedback report.
- Portfolio. The portfolio consists of a descriptive and a reflective part. The descriptive part contains documents that demonstrate the professional performance on the seven CanMEDS competences. In the reflective part, consultants document their reflections on their performance. Consultants were recommended to create a portfolio with the information they had at that point in time and to follow up on it in a structured manner. Consultants were asked to submit their portfolio to the facilitator 2 weeks in advance of the assessment.
- Assessment interview. A trained facilitator discusses the portfolio and the feedback report with the consultant.
- Personal development plan. Consultants formulate improvement objectives in their personal development plan.

Figure 1. Elements and procedures of the assessment system



On the basis of an earlier systematic review, we selected two methods of MSF that were proven to be valid and reliable for consultants.[10] This included: the PAR Program developed by Violato et al[11] and the ABIM introduced by Ramsey et al.[12] Both are examples of MSF methods that are based on questionnaires. PAR contains feedback from colleagues (doctors), coworkers (nurses, allied health care professionals and/or administrative staff) and patients. ABIM is composed of feedback from colleagues (doctors) only. For both PAR and ABIM, we gave respondents the

opportunity to provide narrative comments to increase the specificity of MSF as recommended in earlier literature.[14] With narrative comments we actually mean: a more specific explanation of the ratings given and specific suggestions to improve performance. As the Dutch AAI – developed by Geeraerts and Hoofwijk [13] – was introduced in several hospitals in the Netherlands, some hospitals preferred to test this method. AAI is a purely qualitative method: colleagues and coworkers are being asked to mention three strengths and three suggestions for improvement (narrative comments) for the consultant. These narrative comments are collected, summarised and fed back to the consultant by a facilitator (see Table 1 for an overview of the methods).
Methods were nested within hospitals as it was not feasible to randomize (A detailed description of the implementation strategy and the methods used is available from the investigators).

Table 1. Overview of MSF methods

MSF method	Numbers of hospitals and participating consultants	Colleagues [doctors]	Coworkers (nurses, administrative staff other allied health care professionals	Patients
PAR	3 hospitals, 45 consultants	8	8	25
ABIM	2 hospitals, 30 consultants	15	-	-
AAI	3 hospitals, 45 consultants	Selectively approached by facilitator, on average 5.6 respondents		

PAR = Physician Achievement Review, developed by Violato et al; ABIM= American Board of Internal Medicine method, developed by Ramsey et al.; AAI= Appraisal Assessment Instrument, developed by Geeraerts and Hoofwijk

Sample

In the eight participating hospitals, eight hospital project leaders and 42 facilitators were appointed. Hospital project leaders were consultants (six) or quality assurance managers (two). Facilitators had varying backgrounds: 34 had a primary medical degree, and eight held degrees in related disciplines such as psychology (six) and pharmacy (two). From the facilitators with a medical degree, 16 were in general medicine (psychiatrists, paediatricians, internists, neurologists, cardiologists, etc.), 13 were surgeons (urology, gynaecology, general surgery, ENT, ophthalmology, orthopaedics, etc.), and five were anaesthesiologists. 35 facilitators (83 percent) attended the training provided; 38 facilitators completed the pilot, four of them pulled out because of lack of time (two facilitators) and personal circumstances (two facilitators). In total, 109 consultants voluntarily participated in the study. They were from varying specialties: general medicine (43), surgery (35), anaesthesiology (10) and diagnostic specialties such as radiology, pathology and microbiology (21); 83 were male and 26 were female. Hospital project leaders ensured that consultants were paired up with facilitators from a different specialty. All facilitators and consultants were approached to participate in the study.

Data collection

The study was a cross-sectional survey study. The evaluation incorporated two surveys to answer our research questions. The study was reviewed by the Institutional Review Board and met the criteria for exemption from further review.

FEASIBILITY

We separated feasibility into two concepts: the time investment required by all people involved to carry out one assessment and the response rates achieved with the MSF-method used. Time investments of different people were collected with (1) facilitator checklist, (2) consultant questionnaire and (3) MSF questionnaires. The facilitator checklist and consultant questionnaire were developed for this study and were subjected to piloting in order to ensure face validity and clarity. Facilitators and consultants were asked to document their time investments in, respectively, a 10-item closed response checklist (facilitators) and a 12-item questionnaire (consultants). Respondents on the MSF were asked how much time they had spent on giving feedback to their colleagues at the end of the MSF questionnaire. Response rates were gathered by a web-based system for PAR and ABIM or through the facilitator (AAI).

TOPICS ADDRESSED

The topics that were addressed in the assessment interview were measured with the facilitator checklist. Facilitators were asked to fill out which CanMEDS competences were addressed in every assessment interview. Non-responders were reminded up till three times through electronic and paper mail.

PERCEIVED IMPACT UPON CLINICAL PRACTICE

Impact upon clinical practice was defined according to a modified version of Kirkpatrick's model introduced by Curran and Fleet.[15] This model identifies four levels of effectiveness: namely, satisfaction (level 1), learning outcomes (level 2), performance improvement (level 3) and patient or health outcomes (level 4). Level 3 (performance improvement) can be further separated into (self)-reported change in performance (level 3A) and a measured change in performance (level 3B). We investigated impact upon clinical practice on level 1, level 2 and level 3A with a questionnaire for consultants. Eight items of the consultant questionnaire concerned the impact upon clinical practice and consultants were asked to fill out the questionnaire after the assessment interview. Items were to be rated on a Likert scale of 1–5 (1=totally disagree, 5=totally agree). We asked for consultants' satisfaction with the assessment system (level 1). In addition, we questioned the number of improvement objectives they formulated (level 2). Furthermore, we measured consultants' perceptions of the impact of MSF on future clinical practice (level 3A). Each questionnaire allowed space at the end for consultants to provide additional free text information about their experiences with the assessment system. Two reminders were sent to non-responders at 4-weekly intervals.

Data analysis

To investigate feasibility, the total time investment was estimated and response rates of PAR, ABIM and AAI were calculated. We used descriptive statistics to analyse the topics addressed in the assessment interviews.

Scales of the consultant questionnaire were constructed on the basis of an exploratory factor analysis. Principal component analysis was performed followed by varimax rotation. We included

only factors with an eigenvalue of >1. Scales were composed on the basis of the results of the factor analysis and a reliability analysis was performed for each scale to identify (and remove) the items with a negative effect on reliability. For further analysis, sum scores were averaged for each scale. Sumscores were log-transformed because of the skewness of data. Descriptive statistics were calculated for the complete set of 12 items of the consultant questionnaire. Because our dataset included individuals (consultants) nested within hospitals, we used generalized estimating equations (GEE) to determine the relationship between perceived impact upon clinical practice and MSF method used. We adjusted for multiple known and hypothesized predictors including gender, age and specialty of the consultant and gender and specialty of the facilitator. We accounted for clustering of consultants within hospitals. Additional GEE analyses that included binomial distributions produced results that were not materially different, so only the primary results are reported. Free text responses on the consultant questionnaire were also analysed.

Results

Response and validity of research questionnaires

We received 82 checklists from facilitators out of 109 assessments (return rate 76 percent) and 89 questionnaires from consultants (return rate 82 percent). The calculated Cronbach's alpha of the consultant questionnaire was 0.86, which supports its internal consistency. Factor analysis (varimax rotation, eigenvalue >1) revealed that two factors explained 69 percent of variance. The two subscales on satisfaction and perceived impact of MSF had high internal consistency reliability (Cronbach's alpha >0.81).

FEASIBILITY. The total time investment required to carry out one assessment appeared to be 8 hours. This consisted of facilitator preparation time (2½ hours), consultant time to compose the portfolio (2½ hours), time to conduct the assessment interview (1 hour) and the total time required for the respondents to provide and submit the feedback for MSF (2 hours). The facilitator preparation time and the time required for MSF varied largely between the methods used (Table 2).

Table 2. Time investment required for PAR, ABIM, AAI in hours (n= 82, return rate= 75%)

	PAR n=27	ABIM n=16	AAI n=39
Time investment facilitator	1 ½ hours	2¾ hours	3½ hours
Time investment consultant	2¼ hours	3 hours	3 hours
Time necessary to provide feedback (colleagues, coworkers, patients)	3¼ hours	2 hours	1 hour
Assessment interview	1 hour	1 hour	1 hour
Total	8 hours	7¾ hours	8½ hours

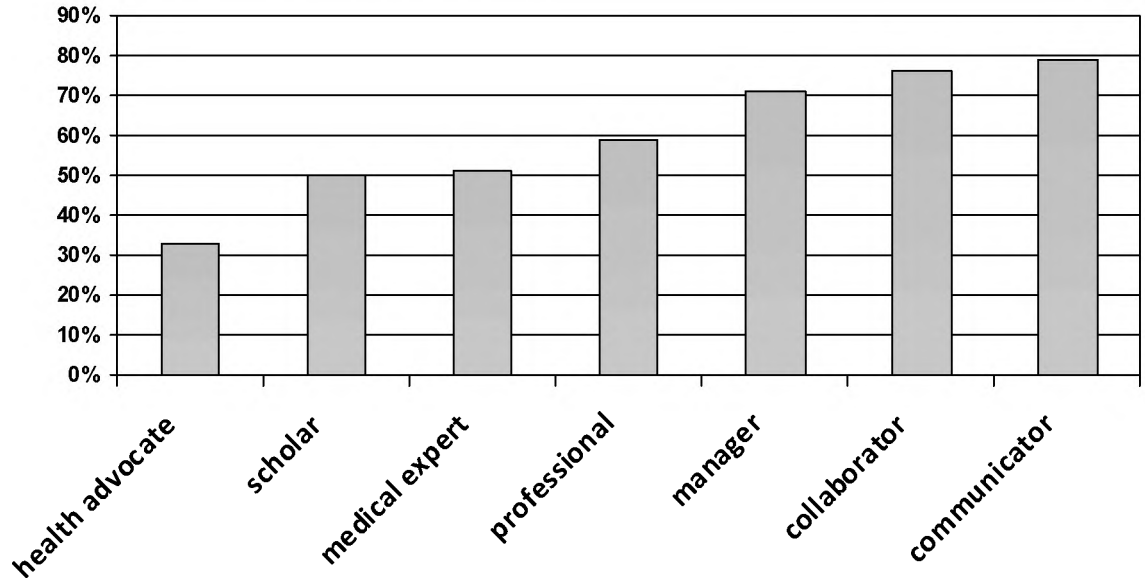
PAR was most convenient in terms of time investment for the facilitator with 1½ hours of preparation time. AAI took significantly more time of the facilitator because he/she had to approach respondents personally and summarise the feedback given. The time required for MSF

was most favourable for AAI with 1 hour compared with 3¼ hours for PAR and 1¾ hours for ABIM, because of the fact that the number of respondents who have been approached is less for AAI (on average 5.6 respondents for AAI compared with 15 respondents for ABIM and 40 respondents for PAR).

The response rates ranged from 64 percent to 87 percent and varied between the different methods and groups of respondents. AAI had a response rate of 87 percent, ABIM achieved a response rate of 64 percent. The response rate for PAR was 67 percent for colleagues and 76 percent for coworkers. The option to provide narrative comments was actually used by a minority of respondents: 42 percent of PAR respondents and 37 percent of ABIM respondents provided narrative comments.

TOPICS ADDRESSED. Analysis of the results of the checklist revealed that the roles of the 'communicator', 'collaborator,' and 'manager' – as defined by the CanMEDS scheme – were the most frequent subjects of conversation between the participating consultants and the collegial facilitator. They were discussed in, respectively, 78, 74 and 71 percent of the interviews. The roles of the health advocate, professional, medical expert and scholar were discussed less frequently (Figure 2).

Figure 2. Topics addressed in assessment interviews



PERCEIVED IMPACT UPON CLINICAL PRACTICE

The results of the consultant questionnaire indicated that, in general, consultants were satisfied with the assessment system and showed a positive attitude towards the assessments: 89 percent of consultants would recommend the performance assessments to colleagues, whereas 74 percent believed that repeating assessments on a regular basis would be useful (Kirkpatrick's level 1; satisfaction) (Table 3). The degree of satisfaction differed between PAR, ABIM and AAI: 89

percent of consultants who received an AAI feedback report believed it was useful to repeat the assessments, compared with 53 percent and 75 percent of consultants for PAR and ABIM.

Table 3. Consultants' perceptions of the impact upon clinical practice

		PAR n=32 mean [SD]	% of agree-ment	ABIM n=20 mean [SD]	% of agree-ment	AAI n=37 mean [SD]	% of agree-ment
Satisfaction	I would recommend the performance assessment project to a colleague	4.2 [0.86]	78%	4.3 [0.87]	85%	4.9 [0.35]	100%
	I think it is useful to repeat performance assessments regularly	3.7 [0.99]	53%	4.0 [0.83]	75%	4.3 [0.67]	89%
	I believe these assessments are an important activity	4.1 [1.04]	75%	4.4 [0.81]	90%	4.9 [0.35]	100%
	I expect my professional performance to improve as a result of the assessment	3.8 [0.83]	53%	3.5 [0.95]	50%	4.2 [0.69]	84%
Perceived impact upon clinical practice	The feedback report that I received has increased my self-insight	3.7 [0.81]	66%	3.2 [0.81]	40%	3.5 [0.88]	61%
	As a result of the feedback report I will improve my professional performance	3.7 [0.82]	59%	3.0 [0.95]	25%	3.6 [0.80]	67%
	The assessment interview has increased my self-insight	3.7 [0.90]	63%	3.6 [0.84]	58%	3.8 [0.55]	76%
	As a result of the assessment interview I will improve my professional performance	3.7 [0.86]	63%	3.6 [1.00]	65%	3.8 [0.80]	73%

Means are based on data from a 5-point Likert scale, with higher scores being more favourable in each category

Adjusting for age, gender, specialty and hospital, the overall model of association between the calculated sum score on satisfaction and method used revealed that consultants are significantly more satisfied with AAI compared with PAR and ABIM (see Table 4, regression coefficient beta=0.114, $p<0.001$). Remarks in the free text sections of the questionnaire may explain this finding. Participants using a questionnaire-based method (three for ABIM, four for PAR) highlighted that the (mainly quantitative) feedback they received was not specific enough to improve the clinical practice. This was different for AAI feedback: facilitators collecting MSF using AAI could ask the colleagues and coworkers they approached to clarify the feedback given.
The average number of formulated improvement objectives was 2,8 and ranged from 0 to 5 (Kirkpatrick's level 2; learning outcomes).

A majority of consultants expressed the intention to change their professional performance as a result of the MSF (Kirkpatrick's level 3A: (self)-reported change in performance). Consultants' perceptions of the impact on future clinical practice differed between the different MSF methods tested. As a result of the AAI feedback report, 66 percent of consultants reported the intention to change compared to 61 percent for PAR and 25 percent for ABIM.
Correcting for all other covariates, the overall model demonstrated that the perceived impact upon clinical practice of the AAI and PAR feedback report significantly exceeds the perceived

impact of feedback produced with ABIM (see Table 4, regression coefficient beta=0.115 for appraisal and regression coefficient beta=0.105 for PAR, p<0.001) The reason expressed by consultants in the free text sections of questionnaires and by project leaders and facilitators during conferences was that ABIM lacked feedback from multiple perspectives, especially feedback from coworkers. According to the consultants, coworkers have a specific and accurate view of their strengths and weaknesses.

Table 4. Parameter estimates and SEs (robust) for sumscore models (generalized estimating equations)

	Sumscore general satisfaction Estimated coefficient [SE]	P-value	Sumscore perceived impact feedback Estimated coefficient [SE]	P-value
Method used				
ABIM	Reference		Reference	
PAR	-0.012 [0.039]	0.767	0.105 [0.020]	<0.001*
AAI	0.114 [0.031]	<0.001*	0.115 [0.028]	<0.001*
Age				
> 55 years	Reference		Reference	
<45 years	0.009 [0.048]	0.854	0.092 [0.082]	0.260
46-55 years	0.008 [0.028]	0.787	0.020 [0.062]	0.742
Gender consultant				
Male	Reference		Reference	
Female	0.039 [0.021]	0.068	-0.033 [0.021]	0.111
Specialty consultant				
Diagnostic specialty	Reference		Reference	
Surgery	0.058 [0.021]	0.006*	0.009 [0.040]	0.827
Internal medicine	0.051 [0.038]	0.178	0.030 [0.052]	0.566
Gender facilitator				
Male	Reference		Reference	
Female	0.004 [0.021]	0.849	-0.007 [0.026]	0.798
Specialty facilitator				
Psychology	Reference		Reference	
Surgery	0.061 [0.055]	0.267	0.108 [0.073]	0.139
Internal medicine	-0.039 [0.044]	0.374	0.047 [0.032]	0.139
Diagnostic specialty	-0.011 [0.054]	0.832	0.024 [0.044]	0.586

* p <0.05

Discussion

Summary of main findings

To reflect changes in practice, doctors are required to update their competences throughout their careers. Performance assessments can both evaluate and foster the development of competence. This study shows that performance assessments based on MSF combined with a portfolio and a supportive assessment interview seems to be feasible in hospital settings. It appears that a majority of consultants are satisfied with the assessments. The results indicate that two-thirds of consultants believe that performance assessment will improve their professional performance. This confirms similar findings in other studies concerning MSF.[16,17]

This article adds to earlier research on the use of MSF in three ways. First, the results show that consultants are more satisfied with feedback that includes narrative comments. Second, the

comparison of the three methods highlights the need to include feedback from coworkers (nurses, administrative personnel and/or allied health care professionals) as it increases the perceived impact of MSF on future clinical practice. It appears that by using MSF different professions can bring unique values about others. This has also been shown to be one of the important mechanisms that influence the delivery of interprofessional education.[18] Third, performance assessments of consultants highlight the attention for the competences of communication, collaboration and management. We assume that this focus is representative of the challenges that doctors are facing in delivering patient care in an increasingly complex hospital environment. It may also indicate that the high level of specialization in today's medicine has increased the threshold to discuss medical knowledge and skills with a facilitator from a different specialty. This is suggested by the findings of a UK study which reported that clinical skills and medical knowledge were found to be the main topics of discussion when general practitioners interview each other.[19]

Limitations

The main limitation of this study concerns the voluntary participation of consultants and the fact that returning a questionnaire reflects in itself a certain level of satisfaction which might have contributed to the positive nature of our findings. Second, the survey relied on self-reporting by doctors which was not triangulated with other data. This means that we have only measured consultants' intentions to change future clinical practice and we did not investigate the real impact upon clinical practice. Third, the concept feasibility was restricted to time investments and response rates. Readers should note that we did not measure other aspects of feasibility such as costs and the practicality in larger non-voluntary settings.

Finally, our study was restricted to one health care system and the results may therefore not be extrapolated to other systems. However, we believe our results might have a broader meaning as the reformed Dutch health care system is often cited as an example for others including the United States.[20]

Implications for research and practice

To improve performance assessments, there are a number of considerations. First, to increase the number and value of narrative comments, respondents should be asked more explicitly to provide narrative comments, including specific examples and concrete tips for improvement. Second, policymakers designing a performance assessment system should include the perspectives of coworkers as our study shows their perspectives increase the perceived impact of MSF. The results of our study point to a crucial need to explore if doctors really improve their performance. Although we incorporated a reflection phase and took care of accurate feedback procedures, the intention to use feedback is comparable to earlier studies that lacked a reflection phase and an interview with a skilled facilitator.[16,17] Possibly, other determinants besides reflection and coaching play a role in the impact of performance assessments. To create a natural

fit between continuous doctor learning in practice and quality improvement of health care, future studies should explore if and under which conditions performance assessments do help doctors.

References

1. Davis MH, Harden RM. Competency-based assessment: Making it a reality. Med Teach 2003;25:565–8.
2. Van der Vleuten CPM, Schuwirth LW. Assessing professional competence: From methods to programmes. Med Educ 2005;39:309–17.
3. Cassel CK, Leatherman S, Black C, Clough C, Gilmore IT, Armitage M. Physicians' assessment and competence: USA and UK. Lancet 2006;368:1557–9.
4. Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. Med Teach 2006;28:e185–e191.
5. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 2007;29:855–871.
6. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119:254–84.
7. Sargeant J, Mann K, Sinclair D, Ferrier S, Muirhead P, van der Vleuten C, Metsemakers J. Learning in practice: Experiences and perceptions of high-scoring physicians. Acad Med 2006;81:655–60.
8. Seifert CF, Yukl G, McDonald RA. Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. J Appl Psychol 2003;88:561–9.
9. Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Challenges in multisource feedback: Intended and unintended outcomes. Med Educ 2007;41:583–91.
10. Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim HC, Grol RP. Doctor performance assessment in daily practise: Does it help doctors or not? A systematic review. Med Educ 2007;41:1039–49.
11. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med 1997;72:S82–S84.
12. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655–60.
13. Geeraerts GAG, Hoofwijk HA. Assessing Medical Professionals [in Dutch]. Houten: Bohn Stafleu van Loghum, 2006.
14. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: Perceptions of credibility and usefulness. Med Educ 2005;39:497–504.
15. Curran VR, Fleet L. A review of evaluation outcomes of web-based continuing medical education. Med Educ 2005;39:561–7.
16. Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: The effect of individual feedback. Acad Med 1999;74:702–14.
17. Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med 2002;77:S64–S66.
18. Hammick M, Freeth D, Koppel I, Reeves S, Barr H. A best evidence systematic review of interprofessional education: BEME Guide no. 9. Med Teach 2007;29:735–51.
19. Colthart I, Cameron N, McKinstry B, Blaney D. What do doctors really think about the relevance and impact of GP appraisal 3 years on? A survey of Scottish GPs. Br J Gen Pract 2008;58:82–7.
20. Seddon N. Is the future Dutch? Lancet 2008;372:103–4.

Chapter 5

Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study

Karlijn Overeem
Hub C. Wollersheim
Erik W. Driessen
Kiki M.J.M.H. Lombarts
Geertje van de Ven
Richard P.T.M. Grol
Onyebuchi A. Arah

Medical Education 2009:43:874–882.

Abstract

Objectives

Delivery of 360-degree feedback is widely used in revalidation programmes. However, little has been done to systematically identify the variables that influence whether or not performance improvement is actually achieved after such assessments. This study aims to explore which factors represent incentives, or disincentives, for consultants to implement suggestions for improvement from 360-degree feedback.

Methods

In 2007, 109 consultants in the Netherlands were assessed using 360-degree feedback and portfolio learning. We carried out a qualitative study using semi-structured interviews with 23 of these consultants, purposively sampled based on gender, hospital, work experience, specialty and views expressed in a previous questionnaire. A grounded theory approach was used to analyse the transcribed tape-recordings.

Results

We identified four groups of factors that can influence consultants' practice improvement after 360-degree feedback: (i) contextual factors related to workload, lack of openness and social support, lack of commitment from hospital management, free-market principles and public distrust; (ii) factors related to feedback; (iii) characteristics of the assessment system, such as facilitators and a portfolio to encourage reflection, concrete improvement goals and annual follow-up interviews, and (iv) individual factors, such as self-efficacy and motivation.

Conclusions

It appears that 360-degree feedback can be a positive force for practice improvement provided certain conditions are met, such as that skilled facilitators are available to encourage reflection, concrete goals are set and follow-up interviews are carried out. This study underscores the fact that hospitals and consultant groups should be aware of the existing lack of openness and absence of constructive feedback. Consultants indicated that sharing personal reflections with colleagues could improve the quality of collegial relationships and heighten the chance of real performance improvement.

Introduction

Doctors are faced with many professional demands, innovations and changes in medical knowledge and techniques, the need to collaborate in larger, often multidisciplinary, teams, and patients who are increasingly knowledgeable about their health and health care. As a consequence, it is important for doctors to ensure and demonstrate that their performance is up to standard. As doctors have been shown to have limited ability to self-assess their performance, external assessments are required for accurate appraisal.[1] External assessments are now well established in revalidation programmes in the UK and Canada.[2] In the past, there has been disagreement as to whether revalidation should aim to enhance professional development or to weed out those who are unfit to practice medicine.[3] The current consensus is that revalidation should do both.[4] There are few studies, however, that have systematically examined the formative aspects of revalidation in terms of its impact on doctors' performance improvement.

One of the methods commonly used to assess doctors' performance is 360-degree feedback.[5,6] Currently, 4444 residency programmes in the USA and all foundation programmes in the UK use 360-degree evaluations to assess residents and fellows. Since 1999, 360-degree feedback has been used for family doctors and surgeons in Canada and internists in the USA. It involves the evaluation of performance on various tasks by, firstly, peers with knowledge of a similar scope of practice, secondly, coworkers from allied health professions and, thirdly, patients. Research by Sargeant et al[7] has shown that 360-degree feedback can be instrumental in improving performance, but its impact may be impaired by doctors' emotional reactions to negative evaluations. Moreover, increased awareness of weaknesses is often not enough to induce behavioural change.[8] The literature suggests, however, that performance improvement can be enhanced by a facilitator who delivers the feedback[9] and by stimulating doctors to reflect on feedback.[10] In this context reflection should be interpreted in the sense of 'letting future behaviour be guided by a systematic and critical analysis of past actions and their consequences'.[11] In a recent study, we found 67 percent of the participating consultants who received 360-degree feedback said they intended to improve their performance.[12] Other studies have reported similar results.[13,14] So far, studies have primarily focused on general practitioners' experiences with receiving 360-degree feedback and their perceptions and reactions towards the feedback itself.[7,10] There has been no rigorous research to explore which factors influence the use of 360-degree feedback for change in future clinical practice in hospitals. The aim of this study was to explore which factors represent incentives, or disincentives, for consultants to implement suggestions for improvement from 360-degree feedback.

Methods

Context of the study

In 2007, eight Dutch hospitals participated in a performance assessment project aimed at improving consultants' performance. Consultants are senior doctors in Dutch hospitals who have successfully completed their residency (also known as specialists or attending physicians in the

USA). All participating consultants received a 360-degree feedback report with information derived from questionnaires completed by colleagues, coworkers and patients, and narrative comments from colleagues and coworkers. The questionnaires were based on translations of two validated instruments, namely the Physician Achievement Review (PAR) programme developed by Violato et al in 1997 and the instrument owned by the American Board of Internal Medicine.[15,16] Questions were to be rated on a 9-point scale. By 'narrative comments', we mean 'a more specific explanation of the ratings given and concrete suggestions to improve performance'. The participating consultants collected evidence concerning their performance in the seven CanMEDS roles (medical expert, communicator, collaborator, scholar, professional, manager, health advocate)[17] in a portfolio and provided written self-reflections on their performance. The portfolio and the 360-degree feedback report were discussed with a trained facilitator (a colleague from a different specialty based in the same hospital). The facilitator (also known as a mentor or coach) helps consultants to interpret the feedback, to critically analyse their performance and to use the feedback to guide future performance. Facilitators were offered 1 day of training which included an explanation of the assessment system, training in basic interview skills and role-plays. The consultants also developed a personal development plan (PDP) including improvement goals derived from the feedback. The process has been described in detail elsewhere.[12]

Study design and participants

In the present study we invited participants in the assessment project to attend an individual face-to-face interview. In order to maximise the richness of the data we used maximum variation sampling. A maximum variation sample is a purposefully selected sample of persons who represent a wide range of extremes related to the phenomenon of interest. The factors we thought to be of influence for the study were: gender, hospital, work experience, specialty and positive and negative views on satisfaction and impact expressed in response to a previous questionnaire.[12] Out of 109 consultants who had participated in our previous performance assessment, we selected 27 consultants who had represented extreme responses on a previous questionnaire, ensuring that they differed in terms of work experience, hospital, specialty and gender. We telephoned this selection of 27 consultants to invite them for a face-to-face interview; 23 consented to participate. Four consultants were unable to take part because of lack of time (two), personal circumstances (one) and unknown reasons (one). The participants included 14 male and nine female consultants from eight hospitals and 13 specialties. Participating consultants had varying backgrounds. Ten came from general medicine (psychiatry, paediatrics, internal medicine, neurology, cardiology, etc.), five were surgeons (urology, gynaecology, general surgery, ear, nose and throat, ophthalmology, orthopaedics, etc.), three were anaesthesiologists and five worked in diagnostic specialties (radiology, pathology, microbiology). The study was given expedited approval by the institutional review board because the participants were not patients.

Individual interviews

We conducted the interviews more than 1 year after the initial assessments to maximise the likelihood that the consultants had initiated changes to improve their practice. Having provided verbal consent, the consultants were interviewed in their offices between April and July 2008. The semi-structured interviews, which lasted approximately 1 hour, addressed the following topics:

- the consultant's reactions to the feedback;
- the consultant's opinions and beliefs about the portfolio, the (role of the) facilitator and the assessment interview;
- improvement goals and the consultant's beliefs and opinions regarding actual performance improvement in practice, and
- the consultant's views regarding factors that promote or impede performance improvement.

The interviewer (KO) encouraged the consultants to speak freely and asked them to illustrate their answers with examples from clinical practice. The consultants received a small fee (equivalent to £35) for their participation.

Analysis

All interviews were tape-recorded and transcribed literally with the consultants' permission. The analysis was based on the principles of grounded theory.[18] Two researchers (KO, GvdV) coded all the interviews independently. Codes were assigned to all issues of interest and were constantly renamed, reorganised and redefined within emerging categories. After coding four interviews, the researchers compared their findings and discussed any differences until they reached consensus. When the first open coding of all the interviews was completed, the next stage of the analysis involved axial coding to identify overarching themes and connections between the themes. The two researchers (KO and GvdV) and one medical education expert (ED) met regularly to discuss the coding and interpretation of the data. Saturation was reached after 12 interviews. However, because of the small volume of data for some categories of information, another 11 transcripts were analysed to ensure comprehensive analysis and coverage of data. The two researchers independently assigned the levels of improvement reported by the participants to four categories based on a model of behavioural change in health care: awareness of a need for improvement (Level 1); acceptance of a need for improvement (Level 2); actual change (Level 3), and maintenance of change (Level 4).[19] We analysed by which factors high levels of change were determined with the help of a cross-case display matrix. Finally, three of the participating consultants were asked to read and comment on the results of the analysis to determine whether the data and conclusions accurately reflected the content of the interviews (member checking).[20] This part of the analysis did not necessitate any changes.

Results

Of the 23 consultants, 11 reported making concrete steps towards performance improvement (Levels 3 and 4).

Two examples of steps taken towards performance improvement were described by: an internist who forced himself to wait for 5 minutes before beginning to speak in multidisciplinary sessions in order to give other people the opportunity to think and speak and a surgeon who went to the emergency department every week for a short visit to ensure she knew the names and faces of the registrars working there.

The other 12 participants had not taken concrete steps (Levels 1 and 2). All the consultants mentioned factors that promoted or impeded change. The four main themes that emerged were: contextual factors; factors related to feedback; characteristics of the assessment system, and individual factors. All factors are summarised in Tables 1–4 and illustrated with quotations from the interviews.

Contextual factors

FACTORS RELATED TO THE HOSPITAL OR CONSULTANT GROUP

Factors relating to the hospital or consultant group were consistently characterised as impediments to change. In the Netherlands the majority of consultants are self-employed and work in a partnership with a group of colleagues. In this paper we refer to partnerships of consultants as 'consultant groups'. The factors identified related to workload, culture in the consultant group and commitment from hospital management. Heavy workload was considered an impediment to the implementation of personal improvement goals. Lack of time interfered with taking action on issues such as collaboration (e.g. writing referral letters on time) and evidence-based practice (e.g. keeping up-to-date with the literature). Relevant aspects related to culture in the consultant group included lack of openness and lack of social support. Half of the consultants believed that sharing their PDPs with colleagues would make it easier to implement them because their colleagues could remind them of their intentions and offer tips and support on implementing change. In reality, however, such sharing did not take place. Lack of commitment from hospital management was mentioned as another impediment to performance improvement. There were many organisational causes for suboptimal performance, such as administrative burden and poor collaboration with nursing staff. Some consultants advocated establishing a feedback loop in which key findings from assessments could be reported to hospital management anonymously and on an aggregate level, so that managers would be able to use this information to support consultants in pursuing improvement goals (Table 1).

FACTORS RELATED TO THE ORGANISATION OF HEALTH CARE AND SOCIETAL FACTORS

Some consultants regarded market forces and health care financing as barriers to performance improvement because of increased emphasis on productivity and heavier workloads. Societal factors such as distrust by patients and the general public were also reported as barriers. The

consultants said that some of their colleagues were not strongly motivated to use feedback to improve clinical practice because they saw 360-degree feedback merely as a means to convince the public that their performance was up to standard. For these consultants, assessment represented a tool with which to boost public confidence rather than an incentive to improve performance (Table 1).

Table 1. Contextual factors

Factors identified and comments	
Factors related to hospital and consultant group	***Factors related to society***
1.Workload (Consultant 4): *'There is less time for all sorts of quality improvement schemes which do nothing for production.'*	1. Market competition and health care financing (Consultant 1): *'There is more pressure on us to be nothing more than production line workers.'*
2. Cultural aspects Lack of openness and lack of social support (Consultant 6): *'People should be more open, it would be helpful if you knew that there was a sort of general consensus about certain problems.'*	2.Public distrust (Consultant 13): *'It's the big fear of any doctor: I am being watched and they are saying how badly I am doing. You should be able to get rid of that taboo.'*
3. Lack of management commitment (Consultant 13): *'It should not be laid at the doctor's door but it should be made a joint effort to try and improve performance in that area.'*	

FACTORS RELATED TO FEEDBACK

Taking action to implement suggestions from feedback was related to the hospital culture and to whether feedback was positive or negative. In general, receiving feedback was valued by consultants. However, in their day-to-day experience, hospital culture did not contribute to making them feel comfortable with giving and receiving feedback on performance. If feedback was given, it was mostly concerned with medical errors and rarely related to interpersonal skills. As a result, consultants thought that 360-degree feedback met a need. Consultants reported that negative feedback was generally difficult to accept, especially when it did not resonate with their self-perceived performance. However, after discussing the feedback with others (their facilitator or a family member, for example), they usually no longer perceived the feedback as problematic (Table 2).

Table 2. Factors related to feedback

Factors identified and comments
1.Hospital culture (Consultant 10): *'It is not easy to give each other this type of feedback. The chance of escalation is higher than the chance of starting a constructive dialogue.'*
2. Negative or positive feedback (Consultant 20): *'That is what I mean when I say illuminating, it is often things that you actually do know or half know. But now they are expressed more clearly by others. And that is an incentive.'*

Characteristics of the assessment system

The consultants indicated that an assessment system would be effective if it encouraged reflection and appropriate action.

(SUPPORTED) REFLECTION

According to the consultants, reflection helped them to see that improvements were needed. Examining their strengths and weaknesses relating to the seven CanMEDS roles in a portfolio gave them insight into the quality of their performance. Because it was unusual for consultants to take a systematic look at communication, collaboration and professionalism, a majority thought that composing a portfolio was 'hard work'. They pointed out that the facilitator should serve as an objective sounding board to help them gauge the accuracy of their reflections. Finally, consultants expected facilitators to encourage them to reflect by exploring with them in detail the reflections in their portfolios and the feedback they received. Facilitators were valued when they paid equal attention to strengths and weaknesses and categorised and summarised the feedback and information in the portfolio to prevent key issues from becoming lost in an overload of detailed information. Consultants indicated that they tended to focus on either their strengths or their weaknesses and they believed that facilitators could counteract this type of 'selective memory' (Table 3).

Table 3. Characteristics of the assessment system

Factors identified and comments	
(Supported) reflection	***Incentives to undertake action***
1.Portfolio (Consultant 9): *'Once you start to think about it explicitly for each domain you begin to see things more clearly.'*	1.Concrete goal setting in personal development plan (Consultant 21): 'Two things were really helpful. For one thing, it was about concrete and achievable things, I think that is really essential.'
2.Facilitator skills: Exploring feedback and reflections in detail (Consultant 1): *'I am convinced that unless people receive some guidance in this they tend to remember mostly what they want to hear.*	2.Facilitator skills: Encouraging specificity of goals (Consultant 22): *'It's a good thing that she [the facilitator] has managed to reduce the issues that need attention to a concrete number of items... and that there aren't any items that are unachievable.'*
3.Facilitator skills: Objectivity (Consultant 12): *'He did that very well and kept an appropriate distance.'*	3. Annual assessments (follow-up) (Consultant 11): *'At a certain point I need to go back to that and then I have to consider: "What have I actually done about that?" And, well, that sort of forces you to actually do it that way.*

INCENTIVES TO UNDERTAKE ACTION

The consultants thought that effective performance assessment stimulated them to take action when it promoted goal setting and included follow-up interviews. Consultants preferred concrete goals to vague intentions and thought facilitators could help them set achievable goals. They also indicated that annual assessments (follow-up) would stimulate them to take action. Repeated exposure to improvement goals and 'knowing that there will be another assessment' was thought to enhance the likelihood of performance improvement (Table 3).

Individual factors

We identified two categories of attitude-related factors that influenced performance improvement: perceived urgency of change (motivation), and belief in one's ability to effect change (self-efficacy).

Although all the consultants had formulated personal improvement goals, they took different views of the urgency of pursuing these goals. Some consultants regarded their goals as intentions and as 'not very important to achieve' because their performance assessment was generally satisfactory. Other consultants considered themselves unable to achieve their goals (lack of self-efficacy). These consultants indicated that the assessment had frustrated them because they realised that improvement was needed but they had no idea how to achieve it. This was problematic for several consultants and caused negative feelings associated with a sense of not being 'in control' (Table 4).

Table 4. Individual factors

Factors identified and comments
1.Perceived urgency to change (motivation) (Consultant 8): '*I am not going to commit myself to spending so many hours every Thursday night to keep up with my reading. No, I wouldn't go so far, after all it isn't all that important, is it?*'
2.Belief in ability to change (self-efficacy) (Consultant 2): '*No, I think improving that, that is just totally impossible. And also, I think I have done everything in my power, I really have.*'

Interaction of factors

Findings about consultants' notions concerning contextual barriers to change seemed surprising in light of the improvements reported by 11 participants.

This issue was explored in the interviews. The analysis of consultants' narratives suggested that specific facilitator skills (encouraging reflection and specificity of goals) and concrete goal setting might overcome negative contextual factors and were key to performance improvement. All consultants who attained higher levels of improvement mentioned these facilitator skills in relation to encouragement of reflection or goal setting, or they emphasised the importance of concrete and achievable goals. The consultants who did not change mentioned these issues only twice in 12 interviews.

Discussion

In view of the increased prominence of performance assessment in relation to revalidation of doctors, we conducted a qualitative study to investigate consultants' responses to 360-degree feedback and their perspectives on factors they considered critical to the achievement of actual improvement in clinical practice.

Our study demonstrates that, despite negative effects from contextual factors, such as high workload, the financing and organisation of health care and public distrust, 360-degree feedback can lead to progress when facilitators help doctors to handle the feedback and reflection is

stimulated. However, our study also reveals that most consultants experience barriers to improvement, mostly as a result of the failure of hospitals to create a climate that is conducive to collegial support and lifelong reflective learning.

Strengths and weaknesses of the study

A limitation of this study is that the participants were all volunteers. Thus we cannot rule out bias arising from the possibility that we may have examined a group of unusually motivated doctors. Moreover, we cannot exclude the possibility that the non-responders would have reported more or different barriers to performance improvement. However, given that half of the consultants had taken no steps to improve performance, we are fairly certain that we have captured most of the impediments. Secondly, this study relied on self-reporting by doctors on whether they had improved their performance and these self-reported data were not triangulated with other data. It will be clear that no general conclusions can be drawn about the actual performance improvement. However, the aim of this study was not to investigate whether consultants actually improve, but to explore the incentives and disincentives for change. Finally, because our study was restricted to Dutch consultants working in non-academic hospitals, the outcomes may not be fully transferable to academic medical centres, primary care settings and other groups of doctors, such as senior postgraduate trainees. The fact that the data were analysed by three researchers from different professional backgrounds (one clinical researcher, one non-clinical psychology researcher and one medical educationalist) is expected to have enhanced the validity and reliability of the results.

Comparison with existing literature

The information gathered in our interviews supports conclusions from other research. The impeding factors we found have also been identified in change processes of other behaviours (e.g. guideline adherence).[21,22] Our study also resonates with work by Frankford et al,[23] who recognised that 'as doctors work nowadays in large group practices or hospitals that deploy financial incentives and management techniques to control clinical performance it is inaccurate to assume that doctors learn primarily as individuals and remain professional principally by virtue of their individual character and moral choice'. The contextual factors that emerged from this study underline the assumption that successful reflective learning depends on interactions with work settings and colleagues. The culture in consultant groups, as described by our consultants, is not characterised by openness and a supportive climate. This is in line with findings by Akre et al[24], who reported that, compared with nonhospital doctors, hospital consultants described the communication climate as more competitive and less supportive. The consultants in our study specifically pointed to the potential benefits to be gained from capitalising on the momentum for structured feedback created by the 360-degree assessments. Several authors have highlighted the importance of feedback climate at work. Argyris and Schon[25] emphasised that a culture in which people can learn from one another is very important for learning and coping in the

workplace. Our study suggests that facilitators who encourage consultants to reflect, set concrete goals based on their reflections and take action to achieve these goals are crucial in helping consultants overcome perceived barriers to change. These findings are in line with work by researchers in the field of human resource management and education. It has been shown that managers who work with a coach set more specific goals and achieve more improvements than managers who have no coach.[26] In addition, a review of over 100 articles on educational research revealed that goal setting enhanced the use of feedback.[27]

Recommendations for practice and research

We recommend various approaches which we believe may enhance the impact of 360-degree feedback. These approaches should be directed at hospitals and consultant groups and at the assessment system. Our results suggest that we should raise awareness of the existing lack of openness and constructive feedback within hospitals and consultant groups. It would be helpful if consultant groups paid attention to their colleagues' experiences with assessments and discussed their PDPs with them. This may induce group reflection, referred to by Frankford et al[23] as an 'institutionalised process of reflection'. Group reflection can promote cooperative, collegial relationships by enabling consultant groups to analyse different approaches to clinical work and consider the implications of performance feedback.[23] Furthermore, hospital management should recognise that doctors can be stimulated to become lifelong learners and reflective practitioners if the organisation is committed to promoting reflection and learning. Obviously, it would be good for hospital managers to be informed of general assessment results, anonymously and on an aggregate level, because this may catalyse a sense of joint responsibility for ensuring optimal clinical performance.

Vital elements of 360-degree assessments in relation to performance improvement include the provision of trained facilitators, concrete goal setting and follow-up interviews. When they are trained for this role, facilitators should be taught how to promote reflection by exploring feedback in detail and how to motivate consultants to take action by asking them to specify concrete goals for improvement.

This study raises new research questions. Although consultants' views of 360-degree assessment are important, other stakeholders may provide additional meaningful information to understand and guide the feedback process. Questions raised by this study include:

- How and when do facilitators encourage reflection?
- How can the feedback best be processed to encourage improvement?

We are currently studying a group of facilitators to explore these questions. Differences between hospitals and primary care settings should also be studied further. Finally, improved conceptualisation of the existing hospital culture by medical professionals is another important area for further research.

References

1 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence – a systematic review. JAMA 2006;296:1094–102.

2 Dauphinee WD. Self-regulation must be made to work. BMJ 2005;330:1385–7.

3 Norcini JJ. Where next with revalidation? BMJ 2005;330:1458–9.

4 Irvine D. Patients, professionalism, and revalidation. BMJ 2005;330:1265–8.

5 Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim HC, Grol RPTM. Doctor performance assessment in daily practice: does it help doctors or not? A systematic review. Med Educ 2007;41:1039–49.

6 Violato C, Lockyer JM, Fidler H. Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. Med Educ 2008;42:1007–13.

7 Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multi-source feedback: perceptions of credibility and usefulness. Med Educ 2005;39:497–504.

8 Grol R. Personal paper: beliefs and evidence in changing clinical practice. BMJ 1997;315:418–21.

9 Seifert CF, Yukl G, McDonald RA. Effects of multisource feedback and a feedback facilitator on the influence behaviour of managers toward subordinates. J Appl Psychol 2003;88:561–9.

10 Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract 2008;13:275–88.

11 Driessen E, van Tartwijk T, Dornan T. The self-critical doctor: helping students become more reflective. BMJ 2008;336:827–30.

12 Overeem K, Lombarts MJMH, Arah OA, Klazinga NS, Grol RPTM, Wollersheim HC. Three methods of multi-source feedback compared. A plea for narrative comments and coworkers' perspectives. Med Teach 2010;32: 141-7.

13 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med 2002;77(Suppl 10):64–6.

14 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. Acad Med 1999;74:702–14.

15 Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med 1997;72 (10 Suppl):82–4.

16 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655–60.

17 Frank JR. The CanMEDS 2005 Physician Competency Framework: Better Standards, Better Physicians, Better Care. Ottawa, ON: Royal College of Physicians and Surgeons of Canada 2005.

18 Kennedy TJ, Lingard LA. Making sense of grounded theory in medical education. Med Educ 2006;40:101–8.

19 Grol R, Wensing M, Eccles M, eds. Improving Patient Care. The Implementation of Change in Clinical Practice. Maarssen: Elsevier, 2004:67–87.

20 Lincoln YS, Yuba G, eds. Naturalistic Inquiry. Newbury Park: Sage Publications, 1985:357–82.

21 Grol R, Dalhuijsen J, Thomas S, in 't Veld C, Rutten G, Mokkink H. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. BMJ 1998;317:858–61.

22 Kennedy T, Regehr G, Rosenfield J, Roberts SW, Lingard L. Exploring the gap between knowledge and behaviour: a qualitative study of clinician action following an educational intervention. Acad Med 2004;79:386–93.

23 Frankford DM, Patterson MA, Konrad TR. Transforming practice organisations to foster lifelong learning and commitment to medical professionalism. Acad Med 2000;75:708–17.

24 Akre V, Falkum E, Hoftvedt BO, Aasland OG. The communication atmosphere between physician colleagues: competitive perfectionism or supportive dialogue? A Norwegian study. Soc Sci Med 1997;44:519–26.

25 Argyris C, Schon DA. Organizational Learning: a Theory of Action Perspective. San Francisco: Addisson Wesley Publishing, 1978:2–19.

26 Smither JW, London M, Flautt R, Vargas Y, Kucine I. Can working with an executive coach improve multi-source feedback ratings over time? A quasi experimental field study Pers Psychol 2003;56:23–44.

27 Shute VJ. Focus on formative feedback. Rev Educ Res 2008;78:153–89.

Chapter 6

Peer mentoring in doctor performance assessment: strategies, obstacles and benefits

Karlijn Overeem
Erik W. Driessen
Onyebuchi A. Arah
Kiki M.J.M.H. Lombarts
Hub C. Wollersheim
Richard P.T.M. Grol

Medical Education 2010:44:140–147.

Abstract

Context

Mentors are increasingly involved in doctor performance assessments. Mentoring seems to be a key determinant in achieving the ultimate goal of those assessments, namely, improving doctor performance. Little is known, however, about how mentors perceive and fulfil this role.

Objective

The aim of this paper is to expand understanding of the role of mentors in performance assessment.

Methods

Thirty-eight mentors undertook formative performance assessments of their peers in a pilot study. A mixed-methods design was used, consisting of a postal survey (n=28) and qualitative interviews with a subset of mentors (n=11). Individual semi-structured interviews were completed and transcripts were analysed by two researchers using a grounded theory approach.

Results

The results of the survey showed that 89 percent of mentors intended to continue in their mentorship role. Interviews revealed that mentors used several strategies in the assessments, including: contrasting and collating information; posing reflective questions, and goal setting. Mentors experienced difficulty in disregarding their views of the doctors evaluated. Some mentors noticed obstacles with specific interview skills such as 'paying attention to their colleagues' strengths' and 'enabling doctors to find their own solutions'. Mentors reported that they and their organisations benefited from the assessments. The perceived benefits included: improved interview skills; increased solidarity, and increased mutual respect.

Conclusions

The study provides insights into what mentors can do to increase the chance that externally derived information is integrated into doctors' self-assessments. Mainly, mentors used strategies aimed at effectively delivering feedback and encouraging reflection. However, we found that mentors who took part in our study appeared to struggle with a number of obstacles related to: time investment; familiarity with the doctor assessed, and the acquiring of specific interview skills.

Introduction

Ensuring that doctors remain clinically competent throughout their careers remains a challenge.[1] As might be expected in a self-regulating profession, doctors bear responsibility for adequately detecting gaps in their own performance and taking proper actions. However, several researchers have highlighted the fact that, for cognitive (information neglect and memory biases) and socio-biological (doctors become adaptive in order to maintain an optimistic look on themselves) reasons, the adequacy of doctors' self-assessments is limited.[2–4] As a result, more externally driven assessments involving, for example, 360-degree evaluations or clinical audits are required.[5,6] Although these assessments vary, they share the underlying goal of making doctors aware of their practice with the ultimate aim of guiding self-directed learning and improving doctor performance.[7] Nevertheless, studies have shown that doctors make few changes in practice in response to external assessments and their self-assessments seem to be stable over time.[8,9]

As performance assessments are relatively new, research into how we can increase doctors' use of performance data is limited. What we do know is that the process of feeding information from assessments back to individual doctors and reflecting on this information appears to be a key determinant in achieving performance improvement.[7,10,11] Research highlights the finding that a coach or mentor is necessary to guide this process.[12,13] Traditionally, the mentor is a trusted and faithful guide for a person who is on a journey of personal, professional and career development.[14] However, different mentoring models and roles exist. In the context of performance assessment, a mentor should be perceived as someone who helps a doctor to interpret feedback and critically analyse his or her work in order to improve future performance.[15] In several countries, such as the USA, the UK and the Netherlands, there are mentors (also known as appraisers or facilitators) who assist in assessment procedures and discuss feedback reports with peers.[16–18] In a previously reported qualitative study, doctors made clear that mentors must encourage reflection, follow-up and goal setting as important conditions for the use of 360-degree feedback for practice improvement.[19] This paper presents further work towards a better understanding of the role of the mentor in order that we can disclose effective mentoring strategies and illuminate important conditions for a mentoring system. In an attempt to meet this challenge, we designed this study to explore the views and experiences of mentors who participate in doctor performance assessments. We specifically investigated how mentors perceive and fulfil their role in performance assessments that combine 360-degree feedback with a portfolio.

Methods

Setting

In 2007, eight hospitals in the Netherlands participated in a performance assessment project. The aim of the project was to develop and evaluate a performance assessment system that would help to improve doctor performance. The assessment system comprised self-assessments

collected in a portfolio and 360-degree feedback from colleagues, coworkers (nurses or allied health care professionals) and patients. Mentors received the feedback report and the portfolio 2 weeks in advance of the assessment interview. Doctors themselves received the 360-degree feedback report from the mentor during the assessment interview. The role of the mentor was to deliver the 360-degree feedback and to encourage reflection in a face-to-face assessment interview. The outcome of this assessment interview was a personal development plan in which doctors formulated their improvement plans. A total of 109 hospital doctors from varying specialties were assessed across the eight hospitals. Thirty-eight mentors from different specialty backgrounds (12 surgeons, 14 internists, five anaesthesiologists, five clinical psychologists and two pharmacists) were appointed. A project leader selected the mentors on the basis of prior experience, interest in quality improvement and qualities as a good communicator. Mentors were offered 1 day of training which included: explanation of the assessment system; goals of the assessment; basic interview skills (active listening), and role-plays. The emphasis in the training was on the assessment system itself and the procedures for confidentiality and objectivity. The participation of doctors and mentors was voluntary and they were not reimbursed for their work. Doctors were matched with mentors from a different specialty to avoid issues regarding familiarity. For feasibility reasons, doctors and mentors from the same hospital were matched. The assessment system used in our study has been described in more detail in a previous article.[18]

Study design

We undertook a cross-sectional, mixed-methods study in two phases from 2007 to 2008 as part of a larger study into doctor performance assessments in the Netherlands. All mentors conducting performance assessments with their peer-colleagues were invited to participate in a survey probing different areas of performance, including training, preparation, satisfaction with the new role and time investments. The survey had two goals; it aimed to generate an overall view of mentors' current opinions, and to select topics for the in-depth interviews as well as mentors to be interviewed. After initial survey analysis, we used maximum variation sampling to select 11 mentors for in-depth interview. A maximum variation sample is a purposefully selected sample of persons who represent a wide range of extremes related to the phenomenon of interest. The factors we thought to be of influence for the study were: gender, specialty, and positive and negative views on satisfaction as expressed in the response to the questionnaire. We telephoned this selection of 11 mentors to invite them for a face-to-face interview. All mentors consented to participate. The interviews were undertaken in order to triangulate information collected in the survey.

Measures

SURVEY STUDY

We measured mentors' perceptions before and after they had conducted the performance assessments with two separate surveys. These questionnaires were developed for this study and were subjected to piloting in order to ensure face validity. After piloting, two items were deleted from the questionnaire and two items required redefinition. The pre-assessment questionnaire consisted of six items measuring preparation and satisfaction about the training. Mentors filled out the questionnaire after the training. The post-assessment questionnaire included seven items measuring satisfaction, time investments and general views about the benefits. Mentors were asked to fill out the questionnaire after the assessment interviews. All questionnaire items were to be rated on a Likert scale of 1–5 (1 = totally disagree, 5 = totally agree). In addition, each questionnaire allowed space at the end for additional free text to capture mentors' experiences. To encourage response, one reminder was sent to non-responders.

INTERVIEWS WITH MENTORS

Interviews took place with a purposive sample of seven male and four female mentors representing a range of specialties and views as expressed in the questionnaire. The interviews, which lasted 45–75 minutes, were conducted at mentors' offices between June and October 2008 by the first author. Semi-structured questions were used as a guide and covered mentors' perceptions of their role and their experiences with the assessments as a whole. All interviews started with a question about what mentors perceived to be the main goals of the assessment interviews. Subsequently, the mentors were asked to reflect upon the following questions:

1 In your perception, what did you do as a mentor to accomplish this goal / these goals?
2 What did you find difficult and why?
3 Did you perceive any benefits?

These topics were raised from the results of the survey. Mentors were encouraged to speak freely and to raise issues important to them. Anonymous processing and analysis of the interviews were guaranteed.

Data analysis

We calculated descriptive statistics for the complete set of items on the questionnaires. We compared the free text responses on the questionnaire with the help of a cross-case display matrix.[20] The interviews were tape-recorded with the participants' permission and transcribed verbatim. Analysis was carried out by hand using grounded theory to look for broad emergent themes. Two researchers (KO, ED) coded all the interviews independently. A cyclical approach was used to add and adapt codes. After coding four interviews, the researchers compared their findings and discussed any differences until consensus was reached. Single passages of text could generate different codes and similar codes were combined. The codes were then categorised into themes which were discussed by two researchers. The accepted coding and themes were used to

analyse the remaining interviews. We stopped interviewing participants at the point when theoretical saturation was achieved. To validate the analysis, we solicited feedback from two mentors (member checking),[21] which led to no adjustments.

Results

Survey study

A total of 27 of the 38 appointed mentors completed the pre-assessment questionnaire (response rate 71 percent). Similarly, 28 of the 38 mentors returned the post-assessment questionnaire (response rate 74 percent). All initial 27 respondents completed the post-assessment questionnaire. One mentor did not attend the training and 'forgot' to complete the pre-assessment questionnaire. Analysis of the non-responders revealed that their gender distribution, age and work experience did not differ from those of responders. Table 1 summarises the results. About 91 percent of the mentors looked forward to facilitating performance assessments. Although they appreciated the training opportunity, mentors perceived the training to be partly insufficient. In the free text comments mentors explained that they believed more role-play related to delivering negative feedback was necessary. As a result, only 45 percent felt sufficiently prepared to perform the assessments (Table 1).

Table 1. Mentors' opinions before and after the assessments

	Strongly disagree n (%)	Disagree n (%)	Neutral n (%)	Agree n (%)	Strongly agree n (%)
Perceptions prior to the assessments					
I understand the goals of the portfolio	0 (0%)	0 (0%)	1 (4%)	22 (81%)	4 (15%)
I feel competent to explain the goals of the assessment to others	0 (0%)	0 (0%)	1 (4%)	23 (85%)	3 (11%)
I feel well prepared to conduct the assessments	0 (0%)	2 (7%)	13 (48%)	11 (41%)	1 (4%)
I learned a lot during the training for mentors	0 (0%)	0 (0%)	4 (15%)	20 (74%)	3 (11%)
I feel competent to manage difficult cases	0 (0%)	1 (4%)	16 (59%)	10 (37%)	0 (0%)
I am looking forward to perform the assessments	0 (0%)	0 (0%)	2 (7%)	21 (78%)	4 (15%)
Perceptions after the assessments					
I learned a lot due to performing the assessment interviews	0 (0%)	1 (4%)	9 (32%)	17 (60%)	1 (4%)
I am satisfied with the way I have performed the assessment	0 (0%)	2 (7%)	11 (39%)	12 (43%)	3 (11%)
I am willing to continue my appointment as a mentor in the future	0 (0%)	1 (4%)	2 (7%)	21 (75%)	4 (14%)
I would recommend a colleague to be a mentor	0 (0%)	2 (7%)	4 (14%)	21 (75%)	1 (4%)
I found the time that I needed to invest for the whole project acceptable	2 (7%)	8 (29%)	10 (36%)	8 (29%)	0 (0%)
I find the time and cost investments are worth the effort considering doctors' benefits	0 (0%)	1 (4%)	9 (32%)	12 (43%)	6 (21%)
I believe that the performance assessments contribute to the professional development of doctors	0 (0%)	0 (0%)	9 (32%)	15 (54%)	4 (14%)

Unless indicated otherwise, each figure in the table indicates the percentage of mentors who chose the corresponding response category.

After the assessments, mentors indicated they were neutral to positive about their own performance as a mentor; 53 percent agreed with the item 'I am satisfied with my own competence as a mentor' and 37 percent were neutral. A majority (89 percent) of mentors reported that they wanted to continue their appointment as a mentor. About 71 percent of the mentors found their time commitments unacceptably high. However, considering the doctors' benefits, a majority (74 percent) found their investment worth the effort.

Interviews

We report the results for the main topics that were discussed in the interviews, which referred to strategies used to ensure that self-assessments resulted in targeted quality improvements, obstacles encountered with the role of the mentor, and benefits observed.

Strategies used

CONTRASTING AND COLLATING INFORMATION

Mentors indicated they collated the doctors' self-assessments in their portfolios with the external feedback from the 360-degree procedure to prepare for the assessment interviews. They looked for similar or contrasting information. In the assessment interviews with the doctors, mentors tried to encourage recognition of the feedback received by the doctor. They did so by confronting the doctor with the similarities or discrepancies between his or her self-assessment and the 360-degree feedback or by simply asking the doctor whether he or she recognised the feedback:

'I did it [giving negative feedback] by looking for similarities. In that case the portfolio was very helpful. I would say, for example: "Yes, you are busy, others can see that too, and they have suggestions how you might improve by doing that or that."' (Mentor 7)

POSING 'REFLECTIVE' QUESTIONS

Mentors explained their role as similar to 'providing a mirror' by emphasising discordance of information and encouraging doctors to think about it themselves. Mentors mentioned that they attempted to ask open questions – especially 'why' questions – and to let doctors draw their own conclusions to encourage reflection:

'Well, by not drawing all sorts of conclusions yourself, but by asking the person who is being evaluated, what they think. "Does it ring a bell?" or "Why do you think that is?"' (Mentor 5)

GOAL SETTING

Mentors reported that they believed it was their responsibility to ensure that concrete and achievable goals were set. Mentors emphasised that they purported to encourage the formulation of achievable goals and to avoid providing simple solutions. In order to achieve this, mentors indicated that they asked consistently about not only what doctors wanted to change, but especially about how they wanted to change (i.e. by asking the 'how question' instead of the 'what question'):

'If people say, "Yes I should work on this," then I ask not only what they are going to improve but also how. "What exactly are you going to do about this?"' (Mentor 9)

Perceived obstacles

The survey data revealed mixed feelings with regard to mentors' preparedness for and satisfaction with their own performance. In the interviews, some mentors explained this was because they had encountered some obstacles. These obstacles were related to familiarity with the doctor they were assessing and the acquiring of new interview skills.

FAMILIARITY WITH THE DOCTOR ASSESSED

Mentors were unanimous in the notion that neutrality was crucial for a good collegial assessment interview. According to the mentors, a certain distance is necessary to encourage reflection and to prevent the assessment interview becoming a 'cosy chat'. In the eyes of mentors, there exists a potential tension between neutrality and familiarity with the doctor evaluated. According to mentors, it is difficult to disregard their own views of the doctor evaluated, which compromises the neutrality of the assessment interview. They mentioned that, prior to the assessment, they tried to consciously erase their image of their colleague's performance:

'You have to be very objective and honest about the information you get, but well, when you have known someone for several years and you see how they work and it is in line with what you think, then it is hard to avoid this prior knowledge completely.' (Mentor 2)

Furthermore, several mentors recognised that familiarity with the doctor who is being evaluated can make the delivery of negative 360-degree feedback difficult for the mentor. In the eyes of some mentors, negative feedback can lead to a perception of the mentor as a harbinger of bad news. As a result, some of the mentors believed that relationships could be influenced because of anxiety about the breaching of confidence and failure to distinguish between the content of the feedback and the messenger:

'As a messenger, you may have to present results that not only are unpleasant for the person concerned, but can also damage your relationship with that person. Obviously, that is not what you want.' (Mentor 1)

This perception was not universally shared by the mentors:

'No, I don't find [delivering negative feedback] difficult because it is not my task to judge someone. At least that's how I see this role, I am here as a mentor for someone who is looking at himself.' (Mentor 9)

INTERVIEW SKILLS

Mentors indicated some difficulties in developing some of the interview skills necessary for carrying out performance assessments. Firstly, mentors noticed that some doctors tend to consider only their weaknesses and that it is therefore necessary for the mentor to explicitly mention strengths before dealing with weaknesses. However, in their experience, it was difficult to do this. Secondly, some mentors observed that the practice of active listening and enabling doctors to find their own solutions is difficult. They argued that this is difficult for them because in many clinical settings they tend to intervene and offer concrete solutions:

'It is quite complicated to stick to the rules because in a conversation, for instance, you easily tend to relate to what someone is telling you, for example by saying: "That's exactly what happens to me in clinic and you might try doing this or that about it."' (Mentor 2)

Benefits for mentors and organisations

A majority of mentors reported in the survey that they wanted to continue their appointment although the amount of time they had been required to invest had been great. In the interviews, participants spoke in greater detail about their satisfaction and argued that they themselves benefited from the assessments in two ways. Firstly, they acquired new interview skills that they could apply in their daily work. Secondly, they learned from the problems that assessee doctors had dealt with, which gave them insights into how to deal with similar situations.

Moreover, mentors were aware of concurrent benefits to the organisation. Most mentors believed assessments contribute to the development of a better working atmosphere in hospitals. They argued that because they try to have an objective position, their prejudices about their colleague doctors disappear. Additionally, they noted that because they are better informed about what their colleagues think and do, solidarity and mutual respect increase:

'I am also convinced that if this was done for all members of staff, in-house relations would benefit from it. I have noticed that because you have to remain objective, your prejudices, for example, towards a certain radiologist disappear. And the doctor gets to know the mentor as a person facilitating a conversation who does not judge.' (Mentor 9)

Discussion

To our knowledge, this is the first study to explore in greater detail the role of the mentor in performance assessments. In an earlier study, we found that a mentor is vital to the success of performance assessments.[19] A major point of agreement between mentors and doctors concerns the importance of reflection and goal setting in the use of 360-degree feedback. Our current study provides some gain in depth of insight related to strategies mentors can use to increase the

chance that a doctor will internalise an external assessment. Interviews revealed that mentors used several strategies to encourage reflection. Strategies included: contrasting and collating information; posing reflective questions, and goal setting. Mentors' perspectives in this study showed similarities with recent findings in the literature based on theoretical discourse. Mentors explained how they 'contrast and collate information' to emphasise discordance of information. Many researchers have underscored the importance of creating an 'aha moment' that integrates high-quality external and internal data as a catalyst for meaningful reflection and change.[3,22] Further, the reflective questions posed by facilitators rely on theoretical assumptions about how one can nurture the concept of 'self-directed assessment seeking', which refers to the process by which doctors take responsibility for looking outward, seeking feedback and information from external sources and using these data to direct performance improvement.[2] The fact that the strategies chosen by mentors to deliver feedback were not discussed in the training for mentors adds to the evidence base for those strategies because mentors reported that they had discovered these strategies by trial and error. Most mentors did not use these strategies in the first assessment interviews and a reasonable proportion of mentors were dissatisfied with their own performance.

A majority of mentors indicated in the survey that they wanted to continue in their mentorship. Mentors explained this was partly because they and their organisations also benefited from the practice. The finding in our study that some mentors expressed giving (negative) feedback as a burden and were afraid it would aggravate intercollegial relationships is of particular interest in the light of perceived benefits for the organisation. The potential burden of providing feedback to a colleague was underlined in earlier studies amongst appraisers in the UK, who expressed their enthusiasm, but stressed the fact that emotional difficulties and tension exist.[16,23,24]

These conflicting perceptions highlight how important it is that mentors acquire skills in giving feedback while maintaining clear procedures with regard to familiarity and confidentiality.

Strengths and weaknesses

There are some limitations to this study. Because of its explorative nature and the limited number of mentors involved, the generalisability of our findings may be limited. Secondly, the study sample was too small to evaluate the validity and reliability of the survey questionnaires thoroughly. Thirdly, the mentors and doctors involved were volunteers. Nevertheless, we believe our findings have a broader meaning as we included mentors from multiple institutions and disciplines and we continued interviewing until saturation had been achieved. To our knowledge, this is the first study to demonstrate a theoretical underpinning of what mentors can do to increase the chance that external assessments such as 360-degree feedback are utilised, which represents a highly relevant and so far underexplored area.[25,26]

Conclusions and recommendations

Before appointing a mentor, four issues and conditions should be considered. Firstly, the fact that some mentors had problems with delivering (negative) feedback as well as with interview skills (e.g. active listening) might be related to their own lack of experience in conducting formative assessments as well as the fact that the training did not focus on practising these skills. This lack of experience should be addressed by improved training in which mentors exercise these three strategies. Mentoring strategies formatted as questions that may be of help in the assessment interviews are listed below.

COLLATING AND CONTRASTING INFORMATION

- Which differences and similarities do you recognise between your self-assessment and the assessments by others?
- Do you recognise a pattern between the assessments?

POSING REFLECTIVE QUESTIONS

- Why do you think others give you this feedback?
- When do these things happen?

GOAL SETTING

- What do you want to achieve?
- How do you want to pursue this goal?

Secondly, matching mentors with doctors with whom they do not have a personal or intensive working relationship is also recommended to prevent awkward situations arising as a result of familiarity. Thirdly, opportunities for interaction among mentors should be created to give them the possibility to talk about difficulties in giving (negative) feedback and the assessments in general. Fourthly, incentives for mentors should be considered in order to compensate for their outlay of time and energy and to encourage the building of a high-quality mentoring system.

As for research, there are a series of unanswered questions. Further investigations are needed to establish whether doctors truly internalise external assessments and whether this results in performance improvement. Future studies could investigate whether suggestions for improvements presented in 360-degree feedback result in adequate improvement plans. Secondly, the influence of a mentor on the discrepancy between self-assessment and external assessment also deserves further study.

References

1 Hays RB, Davies HA, Beard JD, Caldon LJ, Farmer EA, Finucane PM, McCrorie P, Newble DI, Schuwirth LW, Sibbald GR. Selecting performance assessment methods for experienced doctors. Med Educ 2002;36:910–7.

2 Eva KW, Regehr G. ''I'll never play professional football'' and other fallacies of self-assessment. J Contin Educ Health Prof 2008;28:14–9.

3 Epstein RM, Siegel DJ, Silberman J. Self-monitoring in clinical practice: a challenge for medical educators. J Contin Educ Health Prof 2008;28:5–13.

4 Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice doctors who are unskilled and unaware of it. Acad Med 2001;76(Suppl 10):87–9.

5 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of doctor selfassessment compared with observed measures of competence – a systematic review. JAMA 2006;296:1094–102.

6 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess doctor skills across specialties. Acad Med 2004;79(Suppl 10):5–8.

7 Teleki SS, Shaw R, Damberg CL, McGlynn EA. Providing Performance Feedback to Individual Doctors: Current Practice and Emerging Lessons. RAND Health. [Working Paper.] 2006. http://www.rand.org/pubs/working_papers/WR381/. Accessed on 11 December 2008.

8 Lockyer J, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multi-source feedback data. Teach Learn Med 2003;15:168–74.

9 Lockyer JM, Violato C, Fidler HM. What multi-source feedback factors influence doctor self-assessments? A 5-year longitudinal study. Acad Med 2007;82 (Suppl 10):77–80.

10 Sargeant J, Mann K, Sinclair D, Ferrier S, Muirhead P, Van der Vleuten C, Metsemakers J. Learning in practice: experiences and perceptions of high-scoring doctors. Acad Med 2006;81:655–60.

11 Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract 2008;13:275–88.

12 Seifert CF, Yukl G, McDonald RA. Effects of multisource feedback and a feedback facilitator on the influence of behaviour of managers toward subordinates. J Appl Psychol 2003;88:561–9.

13 Luthans F, Peterson SJ. 360-degree feedback with systematic coaching: empirical analysis suggests a winning combination. Hum Resour Manage 2003;42:243–56.

14 Connor MP, Bynoe AG, Redfern N, Pokora J, Clarke J. Developing senior doctors as mentors: a form of continuing professional development. Report of an initiative to develop a network of senior doctors as mentors: 1994–1999. Med Educ 2000;34:747–53.

15 Driessen E, van Tartwijk TJ, Dornan T. The self-critical doctor: helping students become more reflective. BMJ 2008;336:827–30.

16 Lewis M, Elwyn G, Wood F. Appraisal of family doctors: an evaluation study. Br J Gen Pract 2003;53:454–60.

17 Blank LL, Cohen JJ. Feedback improves performance: validating a first principle. Arch Pediat Adol Med 2007;161:103–4.

18 Overeem K, Lombarts MJMH, Arah OA, et al. Three methods of multi-source feedback compared. A plea for narrative comments and co-workers' perspectives. Med Teach 2010;32: 141-7.

19 Miles MB, Huberman M. Qualitative Data Analysis: an Expanded Source Book, 2nd edn. Thousand Oaks: Sage Publications, 1994:172–245.

20 Lincoln YS, Guba G. Naturalistic Inquiry. Newbury Park, CA: Sage Publications 1985;357–82.

21 Overeem K, Wollersheim HC, Driessen E, Lombarts K, Van de Ven G, Grol R, Arah OA. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. Med Educ 2009;43:874–82.

22 Galbraith RM, Hawkins RE, Holmboe ES. Making selfassessment more effective. J Contin Educ Health Prof 2008;28:20–4.

23 McKinstry B, Peacock H, Shaw J. GP experiences of partner and external peer appraisal: a qualitative study. Br J Gen Pract 2005;55:539–43.

24 McKay J, Shepherd A, Bowie P, Lough M. Acceptability and educational impact of a peer feedback model for significant event analysis. Med Educ 2008;42:1210–7.

25 Violato C, Lockyer JM, Fidler H. Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. Med Educ 2008;42:1007–13.

26 Sargeant J, Mann K, van der Vleuten C, Metsemakers J. Directed self-assessment: practice and feedback within a social context. J Contin Educ Health Prof 2008;28:47–54.

Chapter 7

Factors predicting doctors' performance change in response to multisource feedback

Karlijn Overeem
Kiki M.J.M.H. Lombarts
Juliette K. Cruijsberg
Richard P.T.M. Grol
Onyebuchi A. Arah
Hub C. Wollersheim

Submitted.

Abstract

Purpose

Multi-source feedback (MSF) offers doctors feedback on their performance from peers (medical colleagues), coworkers and patients. Researchers increasingly point to the fact that only a small majority of doctors (60-70 percent) benefit from MSF. Building on medical education and social psychology literature, the authors identified several factors that may influence change in response to MSF. Subsequently, they quantitatively studied the factors that advance the use of MSF for practice change.

Method

This observational study was set in 26 non-academic hospitals in the Netherlands. In total, 458 specialists participated in the MSF program. Besides the collation of questionnaires, the Dutch MSF program is composed of a portfolio and a facilitated interview aimed at increasing the acceptance and use of MSF. All specialists who finished a MSF procedure between May 2008 and September 2010 were invited to complete an evaluation form. The dependent variable was self-reported change. Three categories of independent variables (personal characteristics, experiences with the assessments and mean MSF ratings) were included in the analysis. Multivariate regression analysis techniques were used to identify the relation between the independent variables and specialists' reported change in actual practice.

Results

In total, 238 medical specialists (response rate 52 percent) returned an evaluation form and participated in the study. A small majority (55 percent) of specialists reported to have changed their professional performance in one or more aspects in response to MSF. Regression analyses revealed that two variables had the most effect on reported change. Perceived mentor quality positively influenced reported change (regression coefficient beta= 0.527, $p< 0.05$) as did negative scores offered by colleagues. (regression coefficient beta=-0.157, $p<0.05$). The explained variance of these two variables combined was 34 percent.

Conclusions

Perceived quality of mentor supervision and MSF ratings from colleagues seem to be the main motivators for the use of MSF by specialists. These insights could leverage in increasing the use of MSF for practice change by investing in the quality of mentors.

Introduction

The assessment of doctors' professional performance is an important challenge. Nowadays, multisource feedback (MSF) is a central element of these assessments in several countries. Canada was the first country to introduce MSF questionnaires in revalidation programs for doctors.[1] MSF typically involves the completion of questionnaires by a number of colleagues, coworkers and patients –referred to as 'raters'-, whose responses are summarised to identify the doctors' performance strengths or weaknesses. Doctors also complete a self-rating using questions identical to those on the colleague survey so that scores can be compared.[2] Questionnaires used in MSF have now been validated for use across a range of specialties in Canada, Denmark, United Kingdom and the Netherlands.[3,4] Although the validity and reliability of multisource assessments have been examined, little attention has been given to the formative aspects of MSF and its likely value for performance improvement. In general, feedback can be beneficial, neutral, or negative in its impact on future practice. Previous meta-analyses and reviews have established broad agreement on characteristics that are likely to make feedback most effective.[5-7] Feedback should focus on task performance, not on judgments about the recipient's character or personality. In addition, it should be specific and clear, since the interpretation of less specific feedback may frustrate the learner.

The impact of MSF on change in practice (referred to as 'educational impact') has been subject of several studies. Hall et al in Canada found that 83 percent of participants 'contemplated' changing their behaviour[1], but other studies reported fewer people intending to change their behaviour.[8,9] In a later study in Canada, Fidler et al demonstrated that 66 percent of doctors initiated a change for at least one aspect of performance and this was related to lower mean ratings.[10] Lockyer et al revealed that surgeons reported few change in practice in response to feedback data. They found four factors affecting the likelihood of change in response to MSF: age, the time spent reviewing feedback, surgical specialty and the extent to which self-ratings exceed ratings by others.[11]

The educational potential of MSF has been explored by human resources and psychology researchers as well. A study in 2002 showed for instance that women intend to improve more often in response to feedback compared to men.[12] Consistent with studies in medicine, researchers demonstrated that managers who received lower ratings were more likely than others to improve performance.[13,14] In addition, Brett et al observed that overrating (self-rating exceeds MSF ratings by others) limits the use of MSF for future practice.[15] However, if MSF recipients are coached well, discrepancies between self-ratings and ratings from others may catalyse a perceived need for change.[16,17] Miller et al concluded in a recent review that MSF could lead to performance improvement, although individual factors and coaching sessions seem to influence the response.[18] In a previous study we found that mentors who stimulate reflection could increase doctors' performance change.[19] What is lacking is quantitative empirical evidence confirming the factors found in different studies. Furthermore, the factors identified in business settings such as gender and the discrepancy between self and external assessment, need an

evidence base in MSF in the medical profession. On the basis of previous studies, we had the following hypotheses:

1. Personal factors such as higher age, female gender and non-surgical specialty positively affect change in response to MSF.
2. Positive experiences with the MSF assessments related to mentoring and feasibility of webbased service increase change in response to MSF.
3. Lower MSF ratings or a gap between self-ratings and ratings by others positively affect reported change.

In our study, we aimed to answer the following research question: which factors have the strongest impact on specialists' reported change as a result of MSF?

Method

Study context

Twenty-six hospitals participated in the MSF system in the Netherlands. In these 26 hospitals, in total 456 specialists completed the MSF procedure between September 2008 and December 2010. Besides the collation of MSF ratings from colleagues, coworkers and patients, the complete performance assessment system additionally consists of a reflective portfolio and a facilitative interview with a mentor to increase the acceptance of the feedback and its use for practice improvement. Mentors were offered one day of training which included: explanation of the assessment system, goals of the assessment, basic interview skills (active listening) and role-plays. Specialists were matched with mentors from a different specialty based in the same hospital. The MSF-system was launched in 2007 in three hospitals and a pilot study established the feasibility of this system.[20] The MSF process is facilitated electronically by an independent webbased service and is described in detail elsewhere.[20]

Study design and participants

This study was set up as an observational evaluation study based on questionnaires. We invited all 456 participating specialists to take part in the study. Specialists varied in background and work experience. Half a year after specialists had finished their MSF procedure including an interview with the mentor, they were asked to complete a questionnaire measuring the self-reported change they made in practice and their experiences with the mentor and the webbased MSF service. We also asked them for permission to use their MSF ratings (their self-ratings and the ratings from colleagues, coworkers and patients) anonymously for research purposes. One reminder was sent to non-responders after three weeks. The questionnaire consisted of eighteen items on a 5-point Likert scale. Participants had the opportunity to explain their answers in detail at the end of the questionnaire. This study was exempt from ethical approval according to Dutch law. However, we dedicated considerable attention to the interests of our participants.

Measures

DEPENDENT VARIABLE

The dependent variable was self-reported change. We measured self-reported change by asking specialists to rate the item: 'I have changed my professional performance in one or more aspects in the past six months as a result of MSF' on a 5 point Likert scale (1=completely disagree, 5=completely agree).

INDEPENDENT VARIABLES

We measured three groups of independent variables: personal characteristics, experiences with the performance assessments and ratings on the MSF questionnaires.

1. Personal characteristics

a) Gender

b) Specialty.

We categorised specialties according to specialty type into: 1) non-surgical specialties (internal medicine and subspecialties, paediatrics, dermatology, oncology, psychiatry, radiology, anaesthesiology, pathology and neurology), and 2) surgical specialties (surgery, orthopaedic surgery, urology, gynaecology, ophthalmology, otolaryngology, thoracic surgery, vascular surgery, brain surgery).

c) Years of work experience as a registered specialist.

2. Experiences with performance assessments based on MSF

d) Mentor-supervision quality.

From a previous interview study we developed a scale to measure the quality of mentors' supervision in performance assessments. Included items were: preparation of the interview, the degree of increased self-insight and interviewing skills. Responses were invited on a five-point Likert scale. Cronbach's alpha of this scale was 0.80, establishing its internal consistency.

e) Feasibility of the webbased MSF service

This scale was developed from a previous evaluation study. The scale included: the feasibility, helpfulness of the staff and satisfaction with the webbased service. The items used five-point Likert scales ranging from 1 (completely disagree) to 5 (completely agree). Analysis confirmed the internal consistency of this scale with a Cronbach's alpha of 0.72.

3. Feedback ratings on MSF questionnaires

f) Mean MSF ratings from colleagues. We calculated for each specialist a mean score of all colleagues' ratings on the MSF questionnaires.

g) Mean MSF ratings from coworkers. We calculated for each specialist a mean score of all coworkers' ratings on the MSF questionnaires.

h) Mean MSF ratings from patients. We calculated for each specialist a mean score of all patients' ratings on the MSF questionnaires.

i) Self-ratings. We calculated for each specialist a mean score of all self-rated items on the MSF questionnaires.
j) Discrepancy between self-rating and ratings by others. We calculated a mean gap score by subtracting mean ratings between all raters per specialist from the self-ratings by specialists.

Statistical analysis

Descriptive statistics were calculated for the three categories of independent variables. Sum-scores were calculated for the two subscales on mentor supervision quality and feasibility of the webbased service. MSF ratings by colleagues, coworkers and patients from male and female specialists were compared using unpaired t-tests and one-way ANOVA. A p-value < 0.05 was considered significant.

After the initial analyses the independent variables were tested for univariate relationships with self-reported change in order to select the items for the multivariate analysis. The relationship between self-reported change (a score on a 5-point Likert scale and thus considered as a continuous variable) and the dichotomous variables (gender and speciality) was analysed with the Mann–Whitney U test. The correlation between the other variables and self-reported change was analysed with Pearson's correlation. Variables with $p \leq 0.15$ were found to be eligible for multivariate regression analysis. Multiple regression analysis was used to examine which of the independent variables are decisive in doctors' reported change. The specialists being anonymous, we could not correct for the nesting of specialists within hospitals with a multi-level analysis. We selected backward regression as the multiple regression method. The criteria for entry and removal were .05 and .10 respectively, with listwise exclusion of cases. We used SPSS, version 18.0.1 for the statistical analysis.

Results

Study participants

A total of 236 specialists responded to the survey of a possible 452 (52 percent). Seventeen of the non-responders indicated they had lack of time to complete the questionnaire. Because of anonymity issues, other reasons for non-response could not be retrieved. The participants consisted of 144 men (61 percent) and 92 women (39 percent). The percentage of female specialists reflects the whole population of specialists in the Netherlands well.[21] Specialists participating had on average 14 years of work experience.

MSF ratings and self-reported change

The mean gap between the colleagues' ratings and self ratings was -0.21 with a range from -2.34 to 1.81. Of these, 30.3 percent of specialists were over-raters. Female specialists scored themselves significantly lower compared to male specialists on the self-assessment ($t=-3.2$, $p<0.05$). However, scores from colleagues, coworkers and patients revealed no significant differences based on gender of the specialist. Analysis of variances of the mean gap between

colleagues' ratings and self-ratings' revealed that female specialists were significantly more often under-raters (F=3.986, p< 0.05.) compared to male specialists. A small majority (55 percent) of doctors involved believed that they succeeded in improving their performance as a result of the assessments.

Self-reported change and relationship with the independent variables

Univariate analysis using Mann Whitney U tests for the first group of independent variables, gender, specialty and years of work experience, revealed that none of the personal variables, were significantly associated with reported change (see Table 1 for p-values of the correlations).

Table 1. Personal characteristics and correlation with reported change

Domain	Percentage	Mean	SD	p
Gender - male (n=144) -female (n= 92)	 61% 39%	-	-	0.459
Specialty 1. NON-SURGICAL SPECIALTY: dermatology (n= 8), cardiology (n=4), pulmonology (n=5), internal medicine (n=40), psychiatry (n=5), neurology (n=13), paediatrics (n=26), anaesthesiology (n=19), radiology (n=17) and all laboratory specialties such as medical microbiology, pathology and clinical chemistry (n=46). 2. SURGICAL SPECIALTY: general surgery (n= 17), urology (n=5), orthopaedics (n= 6), gynaecology (n=15), ophthalmology (n=2), otolaryngology (n=8)	 71% 29%	-	-	0.403
Years of work experience	-	14.4	8.18	0.427

In the second category, both perceived mentor quality (r=0.565, p <0.01) and feasibility of the webbased service (r=0.169, p <0.01) were positively associated with reported change and therefore eligible for multivariate analysis. (See Table 2 for p-values and correlations)

Table 2. Experiences with performance assessments based on MSF and correlation with reported change

Variable	Mean	SD	Pearsons' Correlation Coefficient	p
Mentor supervision quality scale	3.49	0.85	0.565	0.000*
Feasibility of webbased service	3.09	0.97	0.169	0.006*

*Variables with P≤ 0.15 were included in multivariate analysis

Among the last category of variables including MSF ratings, only the mean ratings of colleagues (r= - 0.195, p <0.01) and the mean of self-ratings (r=-0.179, p <0.01) were significantly correlated with reported change and therefore eligible for multivariate analyses.

Both these latter correlations were negative, which means that higher self ratings (i.e. more positive) and higher MSF ratings by colleagues (i.e. more positive) were associated with less self-reported change. (See Table 3 for correlations and p-values)

Table 3. Scores on MSF and correlation with reported change

Variable	Mean	SD	Pearsons' Correlation Coefficient	p
Mean ratings from colleagues	8.37	0.69	-0.195	0.005*
Mean ratings from coworkers	8.33	0.48	0.020	0.413
Mean ratings from patients	8.12	0.44	-0.013	0.438
Mean of all self-rated items	8.11	0.49	-0.179	0.009*
Discrepancy between self-rating and ratings by colleagues	-0.21	0.62	-0.023	0.384

*Variables with P≤ 0.15 were included in multivariate analysis

After testing for univariate analysis, only the above mentioned four out of the in total nine variables were selected for multivariate analysis. Within the corresponding multivariate analyses with backward selection, two of the four variables were found to predict self-reported change. First, perceived mentor supervision quality (standardized regression coefficient beta: 0.552, $p<0.05$) seems to positively influence the change reported by specialists. Second, the mean MSF ratings by colleagues (standardized regression coefficient beta:-0.152, $p<0.05$) affects reported change directly. The explained variance of these two factors combined was 34 percent. This implies that higher mean MSF ratings (i.e. more positive) by colleagues, made it less likely that a specialist will report change after the MSF assessment. (See table 4 for the results of the regression analyses).

Table 4. Results of multiple regression analysis of four independent variables and reported change

Independent variables	Standardized Beta	T	p
Feasibility webbased service	-0.102	-1.557	0.121
Mean ratings from colleagues	-0.157	-2.414	0.017*
Mean of al self-rated items	0.055	0.851	0.396
Mentor quality	0.527	7.960	0.000*

*= $p<0.05$

Conclusion and discussion

Main findings

Our national survey succeeded in obtaining specialists' views on change in practice as a result of MSF assessments and investigating the association between reported change and different independent variables. With regard to the key ingredients that determine practice improvement in response to MSF, this study is clear in showing that change in practice seems to depend on better quality of the mentor. Furthermore this study shows that specialists who receive lower ratings from their peers (medical colleagues) tend to report more change in practice. As a finding of serendipity, we found that female participants are significantly more often under-raters and score themselves lower compared to their male colleagues.

Comparison with other literature

The importance of mentoring for the use of MSF is supported by earlier work in this domain. In our previous qualitative study on hospital-based assessments we showed that the use of MSF depends on a combination of concrete goals, mentoring and structured follow up.[19] Based on recent literature, we expected other variables such as gender, specialty, work experience and the discrepancy between self-ratings and ratings by others to be influencing factors as well. This was not confirmed by our current study. Presumably, differences in change with MSF amongst Dutch medical specialists are not based on gender or work experience.

The fact that specialists who receive lower MSF ratings from their colleagues, tend to improve more is in line with earlier studies in business settings as well as in medicine.[11,14] However, only a small majority of specialists (55 percent) reported to have changed. In two previous studies, 66 percent of doctors intended to change or reported to having initiated a change.[11,20] There are several possible explanations which may account for the fact that less specialists reported change. First, the MSF feedback reports offered to specialists contain means and standard deviations which were in a relatively narrow range and therefore it was difficult for specialists to identify areas for improvement. This might be caused by the fact that Dutch specialists receive less critical feedback compared to other countries. Second, a comparison with their peer group was not provided in their MSF reports and therefore some specialists may not have considered their ratings as a need for improvement. Third, Dutch doctors might experience less urgency to change compared to doctors in other countries. Our study revealed that MSF ratings by coworkers and patients are not decisive in specialists' change in response to MSF. This is in contrast with a study by Fidler et al in 1999. They showed that those physicians who reported to change received significantly lower mean ratings (i.e. more negative) from patients.[11] In a previous study, we found that specialists are more satisfied with MSF containing feedback from coworkers.[20] Presumably, MSF ratings from colleagues are decisive in making a change in performance, because colleagues provide more often narrative comments to explain their ratings compared to coworkers and patients. However, this hypothesis deserves further study. Finally, the finding that men are more often over-raters compared to their female colleagues is in agreement with earlier findings in human resource studies. Atwater et al found that men tend to overrate themselves more often compared to woman.[22]

Strengths and weaknesses

We consider the findings of this study in the light of potential study strengths and limitations. This study adds to the literature on MSF by moving beyond qualitative research to an empirical analysis of the influence of various factors on doctors' reported change. Strengths of this study are the anonymity of the questionnaire, reducing the likelihood of socially desirable answers as well as the large sample size. The questionnaire being anonymous, specialists' age and more important the hospital and specialist group they are based in were not available for analysis. It

would have been interesting to investigate the effects of various hospitals and specialty groups on reported change as these factors have been found to be important determinants in previous studies.[19] What also would have been an important variable to compare is the effect of various combinations of specialists and mentor based on gender and specialty, as this match probably plays a role in effective mentoring.[23] Furthermore, because of the anonymity of the data from specialists participating in the project, we could not compare responders with non-responders to see if the group of responders was a representative sample. Finally, we measured specialists' perceptions of their reported change only and we did not triangulate this with other findings to check if these changes had taken place.

Importance for future research and practice

Our findings have several implications. The main finding of this study is that the perceived quality of a mentor is the most significant predictor of doctors' reported change. In contrast, it is well known that formal mentoring programmes are scarce and women have more difficulty in finding a mentor than their male colleagues.[24] In light of the increased prominence of underrating in women, this is even more disappointing. What we need are formal mentoring programs in performance assessments. In a previous study, we investigated which strategies mentors use to achieve that doctors integrate external feedback in their self-concepts.[25] An important implication of this study is that mentors should be well-equipped to perform this role and therefore additional training is needed.

Further avenues for future research are clearly signposted from this study. First, studies investigating real change in practice in response to MSF, for example as observed by others, are necessary to verify our findings. For example longitudinal MSF scores can be compared. Furthermore, a more detailed understanding of the mentor-mentee relationship and its effect on self-assessment would also be valuable. For instance, it is not yet clear how often supportive interviews with a mentor should occur for an optimal effect. Additionally, our findings warrant other studies to determine how MSF data can better highlight the need to improve. Presumably, narrative comments play a role in this and this should be further investigated. We join Archer and Miller[10] in advocating for studies over extended periods in which matched groups of doctors are opposed to different interventions. This will require collaboration within and across nations.

References

1 Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. CMAJ 1999;161:52-7.

2 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med 2004;79:S5-S8.

3 Davies H, Archer J, Bateman A, et al. Specialty-specific multi-source feedback: assuring validity, informing training. Med Educ 2008;42:1014-20.

4 Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. Qual Saf Health Care 2008;17:187-93.

5 Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007;77:81-112.

6 Shute VJ. Focus on formative feedback. Rev Educ Res 2008;78:153-89.

7 Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance*: BEME Guide No. 7. Med Teach 2006;28:117-28.

8 Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. Acad Med 2003;78:S429

9 Burford B, Illing J, Kergon C, Morrow G, Livingston M. User perceptions of multi-source feedback tools for junior doctors. Med Educ 2010; 44:165-76.

10 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. Acad Med 1999;74:702–14

11 Lockyer J, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multisource feedback data. Teach Learn Med 2003;15:168-74.

12 Johnson M, Helgeson VS. Sex differences in response to evaluative feedback: A field study. Psychol Women Quart 2002;26:242-51.

13 Reilly RR, Smither JW, Vasilopoulos NL. A longitudinal study of upward feedback. Pers Psychol 1996;49:599-612.

14 Smither JW, London M, Vasilopoulos ML, Reilly RR, Millsap RE, Salvemini N. An Examination of the Effects of An Upward Feedback Program Over Time. Pers Psychol 1995;48:1-34.

15 Brett JF, Atwater LE. 360 degrees feedback: Accuracy, reactions, and perceptions of usefulness. J Appl Psychol 2001;86:930-42.

16 Sargeant J, Mann K, Sinclair D, et al. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract 2008;13:275-88.

17 Ostroff C, Atwater LE, Feinberg BJ. Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. Pers Psychol 2004;57:333-75.

18 Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ 2010;341:c5064.

19 Overeem K, Wollersheim HC, Driessen E, et al. Why doctors do (not) improve their performance after 360-degree feedback: a qualitative study. Med Educ 2009; 43:874-82.

20 Overeem K, Lombarts MJ, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. Med Teach 2010;32:141-7.

21 van der Velden L, Hingstman L, Heiligers P, Hansen J. Increasing number of women in medicine: past, present and future. Ned Tijdschr Geneeskd 2008;152:2165-71.

22 Atwater LE, Waldman DA, Brett JF. Understanding and optimizing multisource feedback. Hum Resource Manage 2002;41:193-208.

23 Sambunjak D, Straus SE, Marusic A. A systematic review of qualitative research on the meaning and characteristics of mentoring in academic medicine. J Gen Intern Med 2010;25:72-8.

24 Stamm M, Buddeberg-Fisscher B. The impact of mentoring during postgraduate training on doctors' career success. Med Educ 2011;45:488-96.

25 Overeem K, Driessen EW, Arah OA, Lombarts MJMH, Wollersheim HC, Grol RPTM. Peer-mentoring in doctor performance assessments. Strategies, obstacles and benefits. Med Educ 2010:44:140-147.

Chapter 8

General Discussion

In this general discussion, the main findings of the studies carried out in this thesis are presented and discussed in the light of relevant and recent literature. The findings are summarised according to three main categories: methods and instruments to assess professional performance, design of a performance assessment system and contextual factors. On the basis of the findings in this thesis, we present a framework which summarises the factors for success of formative performance assessments in doctors' daily practice. Subsequently, the most relevant methodological considerations are reviewed. Finally, we conclude the general discussion with some recommendations for practice and research, finishing with a future research agenda.

Key messages

Initiatives to increase doctors' individual accountability for their performance call for tools to assess performance and provide feedback as a basis for further professional development. Multisource feedback (MSF) can be such a tool and is now being used on a large scale for doctors in the United States, Canada and the United Kingdom. Medical specialists in other countries such as the Netherlands might benefit from this educational method as well. The main results of the various studies lead to the following conclusions:

Methods and instrument to assess doctor performance (Chapters 2 and 3)

- There is a large potential of valid and reliable methods and instruments to assess doctors' performance. The impact of formative assessment on future practice remains hardly known.

- MSF, which implies the collection of feedback on various competences from colleagues, other members of the clinical team and patients is the method most frequently used for the assessment of doctors' professional performance in clinical practice.

- The MSF instruments to assess hospital doctors' performance originally developed in Canada appear to be valid and reliable in the Netherlands after adaptation and translation for the Dutch situation. Only 2 percent of variance in the mean ratings of peers could be attributed to one biasing factor -member of the same specialist group- (Beta=-153, $p<0.01$). Other factors such as length of the working relationship and gender of the peers do not influence the specialists' evaluations. For coworkers, gender had a small effect on ratings given to specialists (Beta=-0.20, $p<0.01$). Gender accounted for 2.2 percent of variance in coworkers' mean ratings. Reliable results are achieved with MSF in the Dutch context with 5 colleagues, 5 coworkers and 11 patient evaluations respectively.

- The self-ratings on MSF by 146 specialists in our sample are not associated with the ratings by peers, coworkers and patients. However, ratings between peers, coworkers and patients were all significantly correlated with each other.

Design of a newly developed doctor performance assessment system in the Netherlands (Chapters 4 and 6)

- MSF combined with portfolio learning and a peer assessment interview is feasible in Dutch clinical practice. The performance assessment system was well accepted by its participants. The perceived impact of MSF including coworkers' perspectives exceeds the impact of feedback that does not include feedback from coworkers. Doctors are significantly more satisfied with MSF that contains narrative comments. The emphasis in the peer assessment interviews is on subjects concerning collaboration, communication and management. Trained mentors, follow up interviews and goal setting are perceived as essential characteristics of any performance assessment system to bring about behavioral change.

- Mentors use different strategies to ensure that external assessments are incorporated into doctors' self-concepts, such as: posing reflective questions and contrasting and collating information. Mentors experience difficulty in disregarding their views of the doctors evaluated and value more training in specific interview skills.

Contextual factors (Chapters 5 and 7)

- Medical specialists perceive various factors to influence their practice improvement after MSF. Those factors can be divided into: 1) contextual factors related to: workload, lack of openness and social support, lack of commitment from hospital management, free-market principles and public distrust; 2) factors related to feedback, such as the content and discrepancy with self-insights; 3) characteristics of the assessment system, such as mentors to encourage reflection, concrete improvement goals, and annual follow-up interviews; 4) individual factors, such as self-efficacy and motivation.

- Peer interviews in which a colleague facilitates the feedback delivery might contribute to increased solidarity and mutual respect in hospitals.

- In our final evaluation with 250 medical specialists we found that a bare majority of specialists (55 percent) reports to change in response to MSF. Regression analyses revealed that two variables had a significant effect on reported change. Perceived mentor quality positively influenced reported change as did negative scores offered by colleagues. The explained variance of these two variables together was 34 percent. Other variables such as gender, work experience and specialty did not appear to influence change in response to MSF.

We will discuss the results in the next paragraph.

Comparison with other literature

Methods and instruments to assess doctor performance

Validity and reliability of methods

Our review showed that there is a large potential of methods and instruments to assess doctors' performance, such as: video observation, direct observation, simulated patients and MSF. MSF is most often used in clinical practice. Van der Vleuten introduced a conceptual model to define the utility of an assessment method. The model derived utility by multiplying classical criteria as reliability and validity as well as educational impact, the acceptability of the method to the stakeholders and the investment required in terms of resources.[1] Published standards identify three sources of validity evidence: content validity, construct validity and criterion validity and two aspects of reliability: internal consistency and stability (inter-rater reliability, or generalisability).[2]

We found that psychometric studies of instruments used for multisource assessment were limited. This is in line with a review by Evans et al who found that the instruments developed for MSF need further assessment of validity before their widespread use is merited.[3] In numerous countries MSF has been implemented quickly on a nationwide basis and evaluation subsequently and understandably focused on reliability and feasibility. In particular, evidence on the criterion and construct validity is lacking. Criterion validity can be analyzed by showing the correlation of an instrument with another measure of the same trait. (ideally a gold standard) Construct validity implies that the instrument measures the 'hypothetical construct'. This can be measured for example by applying the scale to different populations, hypothesizing different results on a certain scale. After the publication of the review by Evans et al four similar studies were conducted regarding the construct validity of MSF questionnaires. Research into MSF instruments for paediatric residents in the UK showed that residents in year 4 of training scored significantly higher than residents in year 2.[4,5] This finding confirms that raters identify an inherent standard setting, thus adding to construct validity.[5] Doctors' overall mean performance scores on MSF in the UK were significantly correlated with the numbers of colleagues with positive comments ($r = 0.35$; $P < 0.0001$) and negative comments ($r = -0.40$; $P = 0.0003$), adding to construct validity.[6] Finally, analyses of the Canadian MSF tools revealed that differences in factors' scores emerged between specialties.[7] For example, communication was the first factor for psychiatrists, whereas patient management was the first factor for paediatricians and internal medicine specialists. These findings add to construct validity as well. Evidence on criterion validity is mainly lacking due to the absence of a gold standard or other good measures. Further criterion validity for the MSF instruments used in Canada, UK and the Netherlands comparing MSF scores with for example other tools such as mini-CEX and the number of adverse events and for example patient complaints should be studied. In our study, we found some evidence for content validity, as we demonstrated that only one biasing factor influenced ratings by peers (i.e. the fact

whether a rater was a colleague from the same specialist group or not). This factor accounted for only two percent of variance, which is supportive construct validity evidence.

Our finding in chapter 3 that self-assessments are not related with external assessments made by peers, coworkers or patients is in line with the current international literature.[8] Our finding justifies the introduction of external assessments for the evaluation of doctors' professional performance. Recently, in continuous professional development, it has been suggested that self-assessment should always be complemented by other assessments, an approach known as directed self-assessment.[9]

Interestingly, we found in chapter 7 that women more often underestimate their own performance compared to men. This is in line with the current MSF literature. Atwater et al found indeed that women tend to underrate themselves compared to men. [10]

Impact of MSF on performance improvement

The results of our literature search revealed that there is no convincing evidence yet that doctor performance assessments lead to improvement in actual performance. Studies investigating the impact of doctors' assessments are usually conducted on small volunteer based samples and measurements are often self-reported changes by doctors. An important explanation for the lack of evidence may be that the interest in evaluating non-technical skills has been increasing since the last ten years and assessment tools have only been implemented in the last five years. We argue that studies with larger populations and a longer follow up period are necessary to find statistically significant differences.

In the past five years, major changes have occurred in the field of doctor performance assessment. In particular, postgraduate medical education is pushed into a period of major reform. Many countries are developing and implementing competency based training programs for residents with an emphasis on workplace learning and workplace based assesment. Workplace based assessments include different assessment methods such as mini-clinical examinations, direct observation of practical skills, portfolios and multisource assessment.[11,12] Nonetheless, results on the educational impact of assessments in graduate medical education cannot be automatically extrapolated to certified specialists. Some new studies on the impact of doctor performance assessment have been performed since we completed the review. Brinkman et al conducted a randomized trial and report observed changes in performance instead of self-reported change. They showed that MSF positively affected communication skills and professional behaviour among 18 paediatric residents who were offered a MSF report compared to 18 residents who did not receive a MSF report. [13] A longitudinal study with 250 doctors receiving MSF showed that upward changes in performance measured by MSF scores occur, however the effect sizes were found to be small to moderate.[14] The results of 11 studies measuring the impact of MSF were analysed in a recent review. Miller and Archer concluded that MSF could lead to performance improvement.[15]

However, they notice that studies show conflicting evidence. Furthermore, it should be noted that the small effects found in the study by Violato could not be directly related to the implementation of MSF. Possibly, the upward changes that occur are due to other factors such as changes in education or other quality improvement projects. Miller and Archer hypothesised that individual factors and the presence of facilitation have an important influence on the magnitude of the response.[15] This hypothesis on the positive influence of facilitation in the educational impact of MSF has been derived from qualitative research amongst family physicians in Canada who indicated that coaching is necessary to handle negative feedback and is supported by the last study in this thesis.[16,17]

Design and implementation of a newly developed doctor performance assessment system in the Netherlands

As can be concluded from our review, research on how to move from valid and reliable instruments to MSF systems with acceptable and effective feedback content and sustainable impact is lacking. Therefore, the design of the Dutch performance assessment system was not only based on empirical findings but also on educational theories. In the literature, suggestions are being made about the positive influence of a mentor and a reflection phase in theoretical articles and qualitative studies.[16,18,19] Based on these suggestions, a portfolio and a mentor were incorporated in our MSF system with the aim to increase the acceptance of feedback.[20,21] The evaluation study in chapter 4 demonstrated that performance assessments based on MSF combined with a portfolio and a supportive assessment interview is feasible in hospital settings. Two-third of specialists believed that performance assessment will improve their professional performance. This is comparable to earlier studies concerning MSF that lacked a reflection phase and a supportive interview with a mentor.[22,23] Nevertheless, we must conclude, it does not make sense to expect that simply incorporating a mentor in a performance assessment system will automatically increase the impact of MSF. We conducted additional studies to elucidate which elements of a MSF system contribute to its success and we will discuss the results of these studies in the next 2 paragraphs.

Narrative comments and coworkers

One important point to be taken from our studies is that MSF should contain coworker feedback and narrative comments as they increase the impact of MSF and doctors' satisfaction respectively. An important argument for the usefulness of coworker feedback has been highlighted by a recent study by Bullock et al.[24] They showed that senior nurses are more likely to give critical feedback than peer doctors. Presumably, doctors in our study perceived the coworker feedback as useful because critical feedback reflects areas for improvement which might serve as an incentive for improvement. Our findings regarding the satisfaction of doctors with narrative comments confirm the results of three other MSF studies. Burford found similar results with junior doctors who preferred textual feedback.[25] Smither and Ferstl et al found that

MSF recipients pay more attention to narrative comments than to quantitative ratings.[26] These findings are not surprising in the light of an earlier qualitative study by Sargeant et al in which family physicians reflected that the feedback reports containing only numerical scores are often inadequate to identify areas for improvement.[16]
Our findings are underlined by a recent overview of the assessment literature by van der Vleuten et al. They concluded that narrative, descriptive and linguistic information is much richer and more appreciated by learners while the assessment literature is associated with quantification, scoring, and averaging.[27]

Mentoring, goal- setting and follow up
One of the insights from the qualitative study in chapter 5 is that formulating concrete goals for change, follow-up interviews, and incentives for reflection provided by mentors advance doctors' use of feedback for practice improvement. The positive effect of defining concrete goals is in line with MSF studies in business settings and quality improvement research. Studies have shown that people who are goal and outcome oriented are more likely to take positive steps towards change.[28,29] Doctors reported that discussing the feedback with a mentor helped them to accept the feedback and use it for change. This is confirmed in studies by psychology researchers who found that an interview in which reflection is stimulated increases self-efficacy.[30] Other studies found that managers who received MSF and worked with a coach improved significantly more than others.[31] In this respect, mentors in our mentor study (Chapter 6) indicated they saw stimulating doctors to internalise external assessments as their main goal. They reported to use several strategies to achieve this goal such as 'posing reflective questions' and 'contrasting and collating information'. However, mentors indicated that they do not always find themselves well equipped to perform this task and would prefer more training. The fact that mentors need more and focused training might explain why doctors' intentions to improve performance do not exceed the intentions measured in systems that lack a mentor and stimulation of reflection. We hypothesise that in the first phase of implementation, the quality of mentors differed because not all mentors received a training. As a result, not all doctors profited from the reflection stimulated by a mentor. This is confirmed in our last study in Chapter 7 in which we established that mentor quality is one of the two significant predictors of doctors' reported change.
Although the need for a mentor to discuss the results of MSF has been suggested before, in many settings – for example Canada- a mentor is lacking, likely due to financial or organizational limitations. In our third study we found indeed that mentorship is time consuming. In total, it took mentors on average 180 minutes to prepare and perform one supportive interview. This is in agreement with previous studies who found that lack of time is an important barrier for mentoring.[32,33]As mentorship is time and cost-consuming, future studies could try to compare the effect of MSF between settings with and without a mentor and unmask how often assessment interviews should take place in order to have a maximum effect.

Contextual factors for performance improvement with MSF

The qualitative study in chapter 5 provides new insights into why MSF does not always meet expectations in hospital settings. To our surprise we found that hospital doctors perceive many contextual factors, such as high workload, cultural aspects such as lack of openness and social support, free-market principles and public distrust as barriers for improvement after MSF.

Some of these factors have also been identified in change processes of other behaviours (e.g. guideline adherence), however we did not expect them to be present in improving professional performance with respect to personal goals such as enhancing communication and collaboration.[34,35] The impact of contextual factors on professional performance improvement still have to be demonstrated quantitatively. However, several studies support their presence, for example regarding the use of public reporting of performance (PR) and pay for performance (P4P) in health care. It has been mentioned in the literature that the use of performance assessment data for other purposes may be detrimental for acceptance of performance assessment by engendering fear and suspicion among doctors.[36] Additionally, the slow embracement of performance assessment by doctors is claimed to be due to the fear of increased litigation risks and public reporting.[37,38]

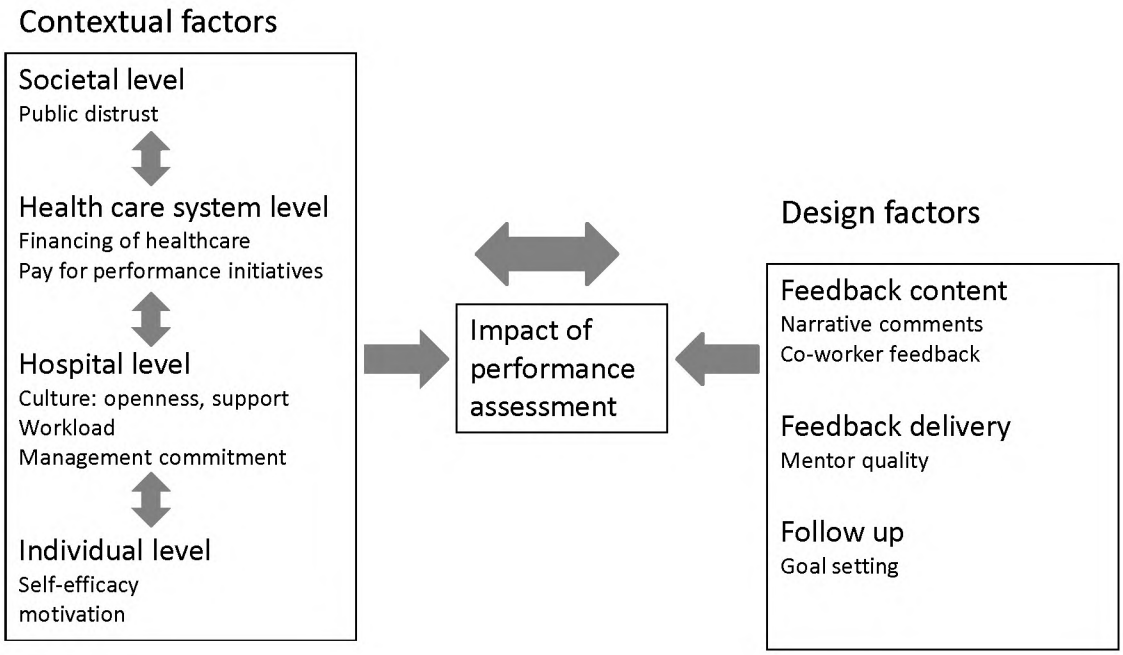
Some specialists indicated that their hospital culture did not contribute to making them feel comfortable with giving and receiving feedback on performance. Others also hypothesised that cultural aspects might play a role in the disappointing results concerning doctors' perceived usefulness and effectiveness of MSF.[39,40] In this respect, we found promising results of mentoring as well. Mentors themselves reported that conducting the assessments increases mutual respect and solidarity. This confirms findings by psychology researchers who have shown that an interview in which reflection is encouraged increases bonding with the interacting partner.[30] Possibly, a mentoring system might contribute to performance improvement with MSF in two ways. First, by increasing acceptance of feedback and self-efficacy of the doctor involved. Secondly, by creating a culture of solidarity and openness in which collaboration flows naturally. By influencing this culture, an important hindering factor –organizational culture– for performance improvement might be tackled. In view of these possible cultural changes, a recent study by Keroack et al in 79 academic medical centres in the United States is promising. They found that a high score in hospitals on patient safety and equity of care was associated with a culture of collaboration.[41] Further exploration of these possible effects is of course necessary.

A new framework for the success of doctor performance assessments

In 2007, Clancy asserted there is a need to help health care move beyond what is easily measured to developing the intellectual and conceptual architecture that will help us close the yawning gap between the promise of health care today and current performance.[42] With this statement Clancy aimed at health care system performance. However, for doctor performance assessments as well, the enhanced capacity to assess performance will not in itself add value to quality of care or result in substantial improvements of that health care. There is lack of evidence on acceptable

feedback delivery and sustainable impact of that assessment. As a consequence, there is a similar need to develop a conceptual architecture that summarises factors influencing the impact of doctor performance assessments on clinical practice. Also in other disciplines there is lack of an evidence base regarding the design of 360-degree assessment systems and how performance feedback can be made acceptable and effective.[43] In 2006, Denisi concluded that 360-degree performance appraisal has been the focus of considerable research for almost a century. Yet, this research has resulted in very few specific recommendations about designing and implementing appraisal and performance management systems whose goal is performance improvement.[44] In this thesis, we succeeded in identifying several factors which support the success of performance assessment. Based on the findings in this thesis, we suggest a comprehensive framework for doctor performance assessment systems. (See Figure 1) The model is based on the two categories of factors previously described in this chapter: system design factors and contextual factors.

Figure 1. Conceptual framework for the impact of doctor performance assessment



In the studies in this thesis we could establish three decisive factors for the impact of doctor performance assessment that were primarily found in the literature in other disciplines: content of feedback, delivery of feedback by mentors and concrete goal setting. We have elucidated several other factors in our studies: contextual factors related to society, health care system and specialist groups and individual factors such as self-efficacy and motivation and follow up.
The need for theoretical frameworks in medical education research has been highlighted recently by several researchers.[45] The integration of concepts from disciplines such as psychology, sociology and human resource management in our framework might lead to fruitful cross-fertilisation.[46] Future studies might benefit from our framework and perform studies in larger cohorts which enable the inclusion of a number of contingent factors. This will demonstrate if the framework will hold up and whether and how it should be changed.

Strengths and limitations
The strengths of this thesis are the methodological rigour and its relevance.

The methodological rigour is reflected in the various research methodologies undertaken and the fact that different stakeholders were included. The multidisciplinary work included: a systematic literature review, qualitative methods, psychometric methods and quantitative methods based on validated and non-validated questionnaire data.
From a psychometrical point of view, our analytical methods to validate the MSF questionnaires were composed of multiple analyses such as exploratory factor analysis, hierarchical regression and generalisability analysis. As for our qualitative studies, we consistently analysed the data with grounded theory, the method described by Miles and Huberman.[47] We used member checking and discussed our findings with an expert group, which supported the trustworthiness of our findings.

A second strength of this thesis is its relevance. Currently, many countries are in search of tools to assess and improve doctors' individual performance. Revalidation programs are going through a period of major reform. The different studies in this thesis illuminated the various factors influencing the success of MSF in ways not published before. This has contributed to a conceptual framework that can inform future researchers and policy makers about the development of assessment programs, a highly relevant subject. Finally, the validity and reliability of the new instruments underlying the Dutch MSF system could be established.

The specific limitations of the studies included in thesis, have been discussed in the separate chapters of this thesis. We will provide general reflections on the limitations of this thesis related to the validity of our findings and the research methods used.

Context

The studies in this thesis were conducted within the context of general (non) teaching hospitals in the Netherlands. As a consequence, caution is warranted for extrapolating findings to other settings such as academic medical centres and other professional groups in the Netherlands and to other countries. We have attempted to overcome these limitations in two ways. First, we performed an extensive literature review. This provided a general overview of findings in different contexts and settings and with different assessment procedures. Second, we compared the findings of all our studies with evidence in other fields such as psychology, human resource management and business studies and other settings such as residency training and primary care.

Participants

A second limitation concerns the largely voluntary nature of the participants in our studies. This group might represent an unusually motivated group of doctors, thus influencing the positive nature of our results. However, it is generally known that a voluntary start cannot be prevented in developmental studies. We used several strategies in order to prevent any undesirable exclusion of participants representing a more negative attitude. For the interviews, maximum variation sampling was used; a technique that ensures the inclusion of participants from different age, gender and attitudes, thus preventing a bias from any single background. Further, we send two reminders to non-responders to achieve a response rate of 82 percent and capture the opinions and attitudes of all people involved.

Methods used and design of the studies

Two studies in this thesis used qualitative research techniques which may be sensitive to bias, due to researchers' unpreventable relative predominance to certain topics in the presented results. Researcher bias happens when the selection of data is influenced by the researcher's preconceptions.[48] We have attempted to rule out researcher bias in several ways. First, two researchers coded all the data independently. Furthermore, we used member checking with several respondents and critical peer review by senior researchers who were not directly involved in the thesis. Some methodological considerations relate to the choice of the measurement instruments. To evaluate doctor performance we selected MSF instruments that were readily available. These MSF instruments had been successfully used in other studies and were validated previously for GPs as well as for medical specialists (surgeons, internists, psychiatrists) in Canada.[7] We chose for this approach for practical reasons as well as for the opportunity to make international comparisons. However, there are some limitations to this pragmatic approach. Firstly, the translation of an instrument from one language or culture to another can be troubled by several difficulties. Language differences present distortions in translations.[49] Secondly, we did not validate the instrument formally before we used them in the setting of Dutch specialists. However, we performed small scale pilot studies in which we

measured acceptation, clarity of terminology and item response to make small adjustments if necessary. We also confirmed content validity in an expert panel.
A final methodological limitation is the fact that our outcome measures relied on self-reporting by doctors. This was not triangulated with other data. Similarly with previous studies regarding MSF, we were not able to do a better job and measure improvement in real practice. There are a few explanations for this. First, some aspects of competence may only change after years. Unfortunately, within the scope of the studies presented in this thesis it was not possible to follow doctors longitudinally for 2 years or more. Second, it is difficult to attribute changes directly to the MSF since aspects of performance may change due to many other influences as well. As a result, strong empirical evidence supporting improvement in routine practice of doctors undergoing assessments is lacking.

Study implications

A variety of recommendations for future practice and research can be derived from the different studies in this thesis. We will address the main stakeholders: doctors themselves, mentors, policymakers and researchers.

Doctors themselves

Participating in assessment and receiving (negative) feedback can be threatening, especially in a field where public reporting and pay for performance initiatives are evolving and there is fear for litigation or damage of reputation. Nevertheless, the embracement of performance assessment by doctors as a means for professional development and quality improvement seems crucial. Changing clinical practice is difficult because it concerns altering sometimes long-established patterns and practices. Also, even after successful change, people may relapse into old routines. To prevent relapse, continuous feedback and reminders are indispensable.[50] For example, doctors can inform people in their immediate working environment about their improvement goals and seek feedback and confirmation.
It has been shown in this thesis that patients and colleagues tend to answer on the upper end of the scale, also known as positive skewness. The interpretation of these scores might lead to limited needs for improvement and perceived impact. Doctors receiving these MSF scores should be convinced that only excellent results are satisfactory. In this respect, doctors also need to have attention for the narrative comments. They are a vital element of effective feedback and add to the educational value of MSF. Doctors involved in MSF as raters should realise that even narrative comments can be too non-specific for doctors to experience learning opportunities. They should try to give concrete feedback including specific examples and tips for improvement.

Mentors

Ensuring that external feedback is integrated in one's self-concept is crucial in the process of practice improvement and change. Mentors have an important but difficult role to fulfil here.

Doctors' emotional reactions can preclude positive effects from feedback.[16] After negative feedback, doctors can pass through phases of shock, anger and denial before being able to accept the feedback.[51] Indeed, systematic reviews show that performance has been noticed to decline following negative feedback.[52] Mentors should guide doctors towards acceptance of feedback. Mentors in our study revealed a lack of experience with particular interview skills and feedback and reflection strategies. Before performing as a mentor, focused training is therefore essential. Mentors need to learn the strategies of contrasting and collating information, posing reflective questions and goal setting as described in chapter 6. The development and implementation of teach the teacher courses- mostly in the context of (graduate) medical education- in which doctors are increasingly trained in these strategies offer a window of opportunity.[53] Additionally, interaction with other mentors to talk about difficulties with giving (negative) feedback and the assessments in general is to be recommended considering the difficulties mentors experienced with delivering negative feedback. The Dutch College of General Practitioners offers different courses for doctors who aim to teach themselves those interview skills. We suggest to offer such courses for medical specialists as well.

Policymakers and hospital managers

As medicine is a self-regulating profession, professional organizations such as the Dutch Organisation of Medical Specialists (Orde van Medisch Specialisten) have a responsibility in ensuring doctors' optimal performance. They do so by the development and implementation of accrediting programs and performance assessment programs based on MSF, described in this thesis. In the Netherlands, doctors and hospital managers –legally– share the responsibility of delivering high quality care. The implementation of MSF introduces a potential area of tension hospital managers and policy makers should be aware of. First, doctors claim ownership of the feedback generated with MSF. In view of the increased public interest in medical errors, the confidentiality of the data generated with MSF is of great importance. The consequences of publicly reporting performance data can be seen in the US where public distrust has undoubtedly increased doctors' vigilance and tendency to avoid potential failure and as a consequence decrease their willingness to participate in performance assessment initiatives.[36,37] Researchers point to the fact that doctors are slow to embrace performance assessment due to the abovementioned issues. On the other hand, this thesis demonstrates that hospitals should support medical specialists with the implementation of performance improvement plans. To prevent issues of doctors' distrust and catalyse a sense of joint responsibility for optimal clinical performance between hospital managers and doctors, hospitals should obviously be informed of general assessment results while ensuring safety and confidentiality regarding the results of performance feedback.[51]

Additionally, specialists experience a lack of support and openness in the current hospital climate. The importance of a culture where people can learn from each other for learning and coping in the workplace has already been emphasised by Argyris and Schon in 1966.[54] Here lies a

task for hospital managers as well as for policymakers. They may encourage and invest in collegial, reflective learning by introducing group reflection and stimulate that personal goals are shared within specialist groups.[55]

Third, with respect to the positive skewness of the results of the questionnaires, presumably the idea of visualizing the outcomes into 'excellent ratings' versus 'sufficient ratings' and 'lower ratings' might visualise deficiencies more clearly. This approach might increase the educational potential of MSF.[56] Recent literature has led us realise that instead of sharpening the instruments – by for example operationalisation of the rating scale –, we should sharpen the people using the instruments.[27] This means that raters should be informed about their task and how to give feedback.

Finally, from our last evaluation study it has become clear that a well-functioning web-based service and feasibility of the methods are important for the success of MSF assessments. Specialists feel they are currently being overwhelmed by evaluations and questionnaires which aim to improve quality. Therefore, new initiatives that aim to combine and link different data on quality of care and doctor performance should be exploited. For instance, one ICT service for reregistration as well as for individual performance assessment can bring together different sources of information and simplify the process. The fact that the MSF questionnaires can be shortened and need to be filled in by five colleagues instead of the eight colleagues in the beginning, offers good future prospects.

Policy implications

As a result of the successful evaluation study described in chapter 4, the Orde recommended medical staff to implement IFMS (Individueel Functioneren Medisch Specialisten) based on the instrument developed by Violato et al[57] or the Appraisal and Assessment instrument developed by Geeraerts and Hoofwijk.[58] In the year 2010, the Dutch Healthcare Inspectorate included participation in IFMS in the primary set of performance indicators for Dutch hospitals. Although this thesis underscores the importance of participation of all medical specialists in IFMS, there are two important comments on this development. First, it should be noted that participation in IFMS as an outcome indicator does not imply that it is successfully implemented. For example, it does not include information about how many mentors were trained and whether the right actions were undertaken to inform all people involved in the feedback procedure. Second, when hospitals' primary goal in IFMS is participation of as many specialists as possible, important issues for the implementation might be overlooked. We found that the most important issues to be improved are: training of mentors, the ICT process and feasibility of the MSF procedure.

Box 1. Implications for practice

- *Narrative comments are indispensable for the educational value of MSF and are recommended to be included in all questionnaires. Assessors in MSF procedures should be stimulated to provide qualitative information.*
- *Mentors should be trained to deliver the feedback and stimulate reflection. Important strategies to be taught are: contrasting and collating information, posing reflective questions and goal setting.*
- *Hospitals need to be informed about the anonymous results of the assessment to catalyze a joint responsibility for optimal clinical performance.*
- *Specialist groups and hospitals should be aware of the positive effect of collegial, reflective learning and are recommended to invest in group reflections.*
- *New initiatives should be exploited that combine and link different data on individual doctors' performance to simplify the process and the webbased service should be improved.*

Researchers

The idea of asking colleagues to assess professional performance, particularly the generic competencies that are less accessible to conventional means of assessments such as clinical examinations remains attractive. However, further research into the validity of instruments and the impact on performance improvement is warranted. For instance, further validity of questionnaires might be sought by exploring the relation between MSF scores and other measures of observed competence, e.g. other quality indicators and the positive and negative comments included in the MSF procedure.

We agree with Miller et al who recently stated that in showing conclusive links between workplace based assessment and performance improvement, a movement towards interventionist, experimental models is necessary.[15] Ambitious research designs which expose matched groups of doctors to different interventions -for example with different procedures for feedback delivery and frequency of stimulated reflection or group reflection- might help in unravelling this link. This requires collaboration within and across health care institutions and nations.

Other important areas for future research are: conceptualizing the culture in hospitals, the feasibility of MSF in other contexts, e.g. primary care and the role of the mentor in integrating external feedback in doctors' self-concepts. Finally, the effect of different contingent factors included in our theoretical framework on the impact of MSF needs to be further studied. Recommendations for future research are summarised in Box 2 on page 116.

Box 2. Future research agenda

- There is a need for further exploration of the validity of the MSF assessments. The relation between MSF scores and other measures of observed competence should be sought.
- Future research should try to unmask the impact of performance assessment on future performance using long term, rigorous designs including experimental models.
- Exploration of different ways to get people act upon externally derived information is necessary
- The culture in hospitals needs to be better conceptualised especially the cultures that support individual doctors to improve performance
- The influence of a mentor on the discrepancy between self-assessment and external assessment deserves further study.
- Larger, longitudinal studies are necessary to disentangle the complex relations between assessment and impact and measure the effect of the various contingent factors summarised in the framework in this thesis
- The feasibility, validity and reliability of MSF in primary care needs to be studied.

6. Final remarks

MSF sees routine doctor practice as an educational endeavour in itself and recognises that the era of the doctor as "a lone ranger" is over.[24] Elsewhere, it has been advocated that -in constructing the links of accountability and education- there is a special need for peer based assessments that target actual performance and meaningful practice outcomes focusing on systems-based quality and personal professional achievement.[25] MSF might very well fulfil that role by building a bridge between quality of care and lifetime medical education. However, MSF will only work and result in professional achievement if adequately managed and if key issues such as credibility and specificity of data, emotional reactions and acceptance, and follow up are taken seriously by all constituent groups.[51] We have some way to go yet.

References

1 van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996;1:41-67.

2 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th Edition. Oxford: Oxford university press, 2008.

3 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. BMJ 2004;328:1240-5.

4 Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in a national programme. Arch Dis Child 2010;95:330-5.

5 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. BMJ 2005;330:1251-3.

6 Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. Med Educ 2009;43:757-66.

7 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med 2004;79:S5-S8.

8 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, f. Accuracy of physician self-assessment compared with observed measures of competence - A systematic review. JAMA 2006;296:1094-1102.

9 Sargeant J, Mann K, van d, V, Metsemakers J. "Directed" self-assessment: practice and feedback within a social context. J Contin Educ Health Prof 2008;28:47-54.

10 Atwater LE, Waldman DA, Brett JF. Understanding and optimizing multisource feedback. Hum Resource Manage 2002;41:193-208.

11 Whitehouse A, Walzman M, Wall D. Pilot study of 360 degrees assessment of personal skills to inform record of in training assessments for senior house officers. Hosp Med 2002;63:172-5.

12 Hrisos S, Illing JC, Burford BC. Portfolio learning for foundation doctors: early feedback on its use in the clinical workplace. Med Educ 2008;42:214-23.

13 Brinkman WB, Geraghty SR, Lanphear BP, et al. Effect of multisource feedback on resident communication skills and professionalism: a randomized controlled trial. Arch Pediatr Adolesc Med 2007;161:44-9.

14 Violato C, Lockyer JM, Fidler H. Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. Med Educ 2008;42:1007-13.

15 Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ 2010;341:c5064.

16 Sargeant J, Mann K, Sinclair D, van d, V, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. Adv Health Sci Educ Theory Pract 2008; 13:275-88.

17 Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. Med Educ 2005;39:497-504.

18 Galbraith RM, et al. Making self-assessment more effective. J Contin Educ Health Prof 2008;28:20-4.

19 Epstein RM, Siegel DJ, Silberman J. Self-monitoring in clinical practice: a challenge for medical educators. J Contin Educ Health Prof 2008;28:5-13.

20 Sargeant J, Mann K, Sinclair D, van d, V, Metsemakers J. Challenges in multisource feedback: intended and unintended outcomes. Med Educ 2007;41:583-91.

21 Seifert CF, Yukl G, McDonald RA. Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. J Appl Psychol 2003;88:561-9.

22 Lipner RS, Blank LL, et al. The value of patient and peer ratings in recertification. Acad Med 2002;77:S64-S66.

23 Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. Acad Med 1999;74:702-14.

24 Bullock AD, Hassell A, Markham WA, Wall DW, Whitehouse AB. How ratings vary by staff group in multi-source feedback assessment of junior doctors. Med Educ 2009;43:516-20.

25 Burford B, Illing J, Kergon C, Morrow G, Livingston M. User perceptions of multi-source feedback tools for junior doctors. Med Educ 2010;44:165-76

26 Smither JW, Walker AG. Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time? J Appl Psychol 2004;89:575-81.

27 Van der Vleuten CPM, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. Best Pract Res Clin Obstet Gynaecol 2010;24:703-19.

28 Roberts GE. Employee performance appraisal system participation: A technique that works. Public Pers Manage 2002;31:333-42.

29 Heslin PA, Latham GP. The effect of upward feedback on managerial behavior. Appl Psychology-Int Rev 2004;53:23-37.
30 Kluger AN, Van DD. Feedback, the various tasks of the doctor, and the feedforward alternative. Med Educ 2010;44:1166-74.
31 Smither JW, London M, Flautt R, Vargas Y, Kucine I. Can working with an executive coach improve multisource feedback ratings over time? A quasi-experimental field study. Pers Psychol 2003;56:23-44.
32 Straus SE, Chatur F, Taylor M. Issues in the mentor-mentee relationship in academic medicine: a qualitative study. Acad Med 2009;84:135-9.
33 Sambunjak D, Straus SE, et al. Mentoring in academic medicine: a systematic review. JAMA 2006;296:1103-15.
34 Grol R, Dalhuijsen J, Thomas S, Veld C, Rutten G, Mokkink H. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. BMJ 1998;317:858-61.
35 Kennedy T, Regehr G, Rosenfield J, et al. Exploring the gap between knowledge and behavior: a qualitative study of clinician action following an educational intervention. Acad Med 2004;79:386-93.
36 Landon BE, Normand SL, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. JAMA 2003;290:1183-9.
37 Donohue SK. Health care quality information liability & privilege. Ann Health Law 2002;11:147-58.
38 Werner RM, Asch DA. The unintended consequences of publicly reporting quality information. JAMA 2005;293:1239-44.
39 Archer JC. State of the science in health professional education: effective feedback. Med Educ 2010;44:101-8.
40 Sargeant J. Multi-Source Feedback for Physician Learning and Change. Dissertation Maastricht: Maastricht University, 2006; 93-118.
41 Keroack MA, Youngberg BJ, Cerese JL, Krsek C, Prellwitz LW, Trevelyan EW. Organizational factors associated with high performance in quality and safety in academic medical centers. Acad Med 2007;82:1178-86.
42 Clancy C. The performance of performance measurement. Health Serv Res 2007;42:1797-1801.
43 Denisi AS, Kluger AN. Feedback effectiveness: Can 360-degree appraisals be improved? Acad Manage Exec 2000;14:129-39.
44 Denisi A, Pritchard RD. Performance Appraisal, Performance Management and Improving Individual Performance: A Motivational Framework. Manag Rev 2006;2:253-77.
45 Bordage G. Conceptual frameworks to illuminate and magnify. Med Educ 2009;43:312-9.
46 Eva KW. The cross-cutting edge: striving for symbiosis between medical education research and related disciplines. Med Educ 2008;42:950-1.
47 Miles MB, Huberman M. Qualitative data analysis: an expanded source book. second ed. London: Saga Publications, 1994.
48 Pope C, Ziebland S, et al. Qualitative research in health care. Analysing qualitative data. BMJ 2000;320:114-6.
49 Hilton A, Skrutkowski M. Translating instruments into other languages: development and testing processes. Cancer Nurs 2002;25:1-7.
50 Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback and physicians' clinical performance*: BEME Guide No. 7. Med Teach 2006;28:117-28.
51 Bracken DW, Timmreck CW, Fleenor JW, Summers L. 360 feedback from another angle. Hum Resource Manage 2001;40:3-20.
52 Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull 1996;119:254-84.
53 Jippes E, Achterkamp MC, Brand PL, Kiewiet DJ, Pols J, van Engelen JM. Disseminating educational innovations in health care practice: training versus social networks. Soc Sci Med 2010;70:1509-17.
54 Argyris C, Schon DA. Organizational learning: A theory action perspective. San Francisco: Addisson Wesley Publishing Company, 1978.
55 Frankford DM, Patterson MA, Konrad TR. Transforming practice organizations to foster lifelong learning and commitment to medical professionalism. Acad Med 2000;75:708-17.
56 Makoul G, Krupat E, Chang CH. Measuring patient views of physician communication skills: development and testing of the Communication Assessment Tool. Patient Educ Couns 2007;67:333-42.
57 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. BMJ 2003;326:546-8.
58 Geeraerts GAG, Hoofwijk, HA. Evaluatie van medische professionals. [in Dutch] Houten: Bohn Stafleu van Loghum, 2006.

Summary

Chapter 1 explains that in the last ten to fifteen years more and more interest has been given to assessing and improving doctor performance. This is necessary because doctors are confronted with knowledgeable patients, ever larger teams and exploding medical knowledge. This requires excellent communication, collaboration and management skills. Skills which are eminently suitable to assess and improve in daily practice. After discussing the changes in health care provision and societal demands that have resulted in the need for doctor performance assessment, Chapter 1 provides an overview of how countries such as the United States, Canada, and United Kingdom have introduced doctor performance assessment systems. It concludes that the current literature does not offer all insights necessary to understand how doctor performance assessment systems can contribute to improving doctor performance and ultimately quality of care. Therefore, this thesis aims to study: how to design and implement a performance assessment system that is valid, reliable, feasible and effective in terms of improving doctor performance?

Chapter 2: Which methods and instruments are available to assess the performance of individual doctors in routine practice and what is known about their feasibility, validity and reliability and impact on routine practice?

In chapter 2, we systematically review the literature in order to provide an overview of methods and instruments available to assess doctor performance. Sixty-four articles met our inclusion criteria. The review succeeded in identifying a large number of methods and instruments. We found six methods that can be used to assess doctor performance: video observation, simulated patients, direct observation, audit of medical records, multisource feedback and portfolio. The methods observed varied greatly in feasibility, reliability, validity and effectiveness. Video observation, direct observation and simulated patients are time consuming methods. It costs on average 2-3 hours to reliably assess one doctor. Multisource feedback (MSF) appeared to be most feasible in terms of time. Reliable results can be achieved with one hour of administrative time. The evidence on the validity of instruments used for MSF was found to be weak. Several studies did not meet established standards of instrument development and omitted essential work on construct and criterion validity. The educational impact of doctor performance assessments is still in its infancy and mainly relies on self-reporting of change. Our literature review revealed that the majority of doctors subjected to MSF or portfolio assessment is satisfied with the evaluation. Between 61 percent and 72 percent (in different studies) of the doctors report a change in their behaviour. The impact of doctor performance assessments on doctor performance and quality of care remains hardly known.

Chapter 3: What are the psychometric properties of three new MSF instruments adapted for the Dutch situation and how many evaluations are needed per specialist for reliable assessments?

Before the widespread use of MSF is merited, it is of vital importance that doctors, managers and patients have confidence in the validity and reliability of instruments applied in MSF. The study in Chapter 3 addresses three aspects of validity and reliability: (1) the initial psychometric properties of three new instruments based on existing MSF instruments, (2) the relationship between the different instruments including self-evaluation, (3) the number of evaluations needed per physicians to establish reliability of assessments. In this observational validation study, 150 specialists from 26 hospitals took part. In total, ratings of 864 peers, 894 coworkers and 1960 patients on MSF were available. Principal components analysis shows that six factors explained 67 percent of variance of the peer questionnaire. The scales comprise: collaboration and self-insight, clinical performance, practice based learning and improvement, coordination and continuity of care, emergency care and time management & responsibility. Principal components analysis of the coworker instrument reveals a 3-factor structure explaining 70 percent of variance. Scales include: relationship with other healthcare professionals, communication with patients and patient care. The principal components analysis of the patient ratings yielded a single factor structure measuring patient-centeredness, accounting for 60 percent of variance in ratings. All instruments appeared to be highly internally consistent with Cronbach's alphas of at least 0.94. Regression analyses revealed that only 2 percent of peer ratings could be attributed to one biasing factor: membership of the same specialist group. Other factors such as gender and age of peers and the length of the relationship with the specialist did not influence the ratings from peers. Across coworkers there was a significant difference in scores on the basis of gender, showing that male coworkers tend to score specialists lower compared to female coworkers. Ratings from peers, coworkers and patients were significantly correlated with each other (Pearsons r= 0.210-0.352, $p<0.001$) but none of them were related with the self-ratings. In contrast to previous studies that demonstrated that eight peer, eight coworker and 25 patient evaluations are necessary, our study revealed that five peer evaluations, five coworker evaluations and 11 patient evaluations are required to achieve reliable results.

Chapter 4: What is the feasibility and perceived impact of a performance assessment system based on MSF and a portfolio in eight Dutch general hospitals and which topics are addressed in the assessments?

On the basis of the results of Chapter 2, we selected two methods: MSF and portfolio learning to evaluate the performance of specialists in the Netherlands. We evaluated its feasibility and the perceived impact by medical specialists before implementing the system nationwide. Chapter 4 evaluates the implementation of this performance assessment system. The performance assessment system was composed of: (i) one of three MSF-methods (namely, Violato's Physician Achievement Review (PAR), the method developed by Ramsey et al for the American Board of

Internal Medicine (ABIM), or the Dutch Appraisal and Assessment Instrument (AAI), (ii) portfolio, (iii) assessment interview with a mentor and (iv) personal development plan. The evaluation consisted of a postal survey for mentors and consultants. The study with 109 specialists and 38 mentors showed that it takes on average 8 hours of time of to carry out one assessment. This consists of: mentor preparation time (2½ hours), specialist time to compose the portfolio (2½ hours), time to conduct the assessment interview (1 hour) and the total time required of respondents to provide and submit the feedback for MSF (2 hours). Analysis of the data reveals that communication, collaboration and management are most frequently the subject of conversation between the participating specialists and the collegial mentor. They are discussed in respectively 78 percent, 74 percent and 71 percent of the interviews. Subjects related to other competencies such as: health advocacy, professionalism and medical knowledge are discussed less frequently. The results demonstrate that 89 percent of specialists completing the evaluation questionnaire would recommend the performance assessments to colleagues. Two third of specialists indicate that they intend to improve their professional performance as a result of the performance assessment. With the help of a regression analysis using generalized estimating equations, it became clear that specialists are significantly more satisfied with MSF that contains narrative comments. (regression coefficient beta =0.114, p-value <0.001) Correcting for all other covariates, the overall model demonstrates that the perceived impact of MSF that includes coworkers' perspectives significantly exceeds the perceived impact of methods not including this perspective. (regression coefficient beta =0.115 for AAI, regression coefficient beta = 0.105 for MSF, p-value < 0.001).

Chapter 5: Which factors are incentives, or disincentives, for doctors to implement suggestions from MSF for improvement?

The aim of the study in Chapter 5 was to explore which factors are incentives, or disincentives, for specialists to implement suggestions for improvement from MSF. We carried out a qualitative study using semi-structured interviews with 23 specialists, purposively sampled based on gender, hospital, work experience, specialty and views expressed in a previous questionnaire. We asked them whether they had improved their performance as a result of the personal development plans they had formulated. We conducted the interviews more than one year after the initial assessments to maximise the chance that the specialists had initiated changes to improve their practice. Transcribed tape recordings of interviews were analysed with grounded theory. The two researchers independently assigned the levels of improvement reported by the participants to four categories based on a model of behavioural change in health care: awareness of a need for improvement (Level 1); acceptance of a need for improvement (Level 2); actual change (Level 3); and maintenance of change (Level 4). Of the 23 specialists interviewed, eleven reported concrete improvements in practice. The other twelve participants reported they had not taken concrete steps. All the specialists mentioned factors that promoted or impeded change. The four main themes of factors that appeared to influence specialists' practice improvement after MSF were:

1) contextual factors related to: workload, lack of openness and social support, lack of commitment from hospital management, free-market principles and public distrust; 2) factors related to feedback; 3) characteristics of the assessment system, such as mentors and a portfolio to encourage reflection, concrete improvement goals, and annual follow-up interviews; 4) individual factors, such as self-efficacy and motivation. Specialists who attained higher levels of improvement mentioned the skills of mentors in relation to encouragement of reflection or goal setting or they emphasised the importance of concrete and achievable goals. We conclude that despite negative effects from contextual factors, such as high workload, the financing and organisation of health care and public distrust, MSF can lead to progress when mentors help doctors to handle feedback and reflection is stimulated. However, our study also reveals that most specialists experience barriers to improvement mostly due to the failure of hospitals to create a climate that is conducive to collegial support and lifelong reflective learning.

Chapter 6: How do mentors perceive and fulfill their role in performance assessments and what do they consider effective strategies for feedback and encouraging reflection?

In addition to specialists' perceptions of change after MSF, Chapter 6 depicts mentors' perceptions of their role in MSF. The aim of the study was to investigate how mentors perceive and fulfil their role in performance assessments that combine MSF with a portfolio. We used a mixed-method design based on 2 postal surveys and semi-structured interviews. The analysis of the data was guided again by the method of grounded theory developed by Miles and Huberman. We invited all mentors in the pilot to participate in a survey probing different subjects such as training, preparation, satisfaction and benefits. The results of the survey show that 89 percent of mentors intended to continue their role as a mentor. The interviews reveal that mentors use different strategies aimed at effectively delivering feedback and encouraging reflection. Strategies include: contrasting and collating information, posing reflective questions and goal setting. Mentors are unanimous in the notion that neutrality is crucial for a good collegial assessment interview. However, we observe that mentors who took part in our study appear to struggle with a number of obstacles related to time investments, familiarity with the doctor assessed and acquiring specific interview skills. An interesting finding is that mentors observed several beneficial effects of conducting the assessments including increased mutual respect and solidarity, and improved interview skills. This finding supports the conclusions in chapter 5 which demonstrates that a mentor is vital for the success of performance assessment.

Chapter 7: Which factor(s) have the strongest impact on specialists' reported change as a result of MSF?

The study in Chapter 7 reports on an observational study based on questionnaire data from 236 medical specialists in 26 Dutch hospitals. This research elaborates on the study in Chapter 5 in which we found that several contextual factors influence the use of MSF and practice improvement after performance assessment. Literature from psychology and human resources

was studied for more theoretical foundation of the study. We found that other factors such as age, gender and the gap between self-assessment and assessment by others influence the use of MSF in other settings. What is lacking, is a study which incorporates those various factors and investigates its impact on reported change in response to MSF with doctors. The study in Chapter 7 addresses this need. Specialists were requested to fill in a questionnaire about their opinions and satisfaction with MSF at least six months after they finished the MSF procedure. We obtained a response rate of 52 percent. The dependent variable was self-reported change in response to MSF. The relationship between three categories of independent variables (personal characteristics, experiences with the assessments and mean ratings) and specialists' reported change were analysed with multivariate regression techniques. A small majority (55 percent) of specialists reports to have changed their professional performance in one or more aspects. Regression analyses indicates that two variables have a significant effect on reported change. Mentor quality positively influences reported change (regression coefficient Beta= 0.527, $p< 0.05$). Scores offered by colleagues are negatively associated with reported change, implying that specialists who received lower scores from their colleagues report more often a change. (regression coefficient Beta= -0.157, $p<0.05$). The explained variance of these two variables together was 34 percent. Interestingly, we found that women's mean self scores are significantly lower than male mean self scores. This study highlights that ratings from colleagues and the quality of mentor are the main motivators for the use of MSF by specialists.

Finally, **Chapter 8** summarises the previous chapters. It provides an interpretation of the studies and compares the results with earlier studies in and outside medicine. The review suggested a wide range of methods to assess doctor performance, however MSF was found to be most feasible in daily practice. We were able to establish the validity and reliability of MSF questionnaires currently used in the Netherlands. Our studies revealed that several factors appear to influence change in practice in response to MSF. The inclusion of feedback from coworkers, narrative comments and a mentor support the use of MSF for change in practice. These findings are in line with studies in other settings. Additionally, we have elucidated several other factors that influence the use of MSF for practice change: contextual factors related to society, health care system and specialist groups and individual factors such as self-efficacy and motivation and follow up. Chapter 8 concludes with a theoretical framework that depicts the various factors responsible for the impact of performance assessments. This framework might be used to guide future studies which enable the inclusion of a number of contingent factors. The chapter ends with a summary of the implications for doctors, mentors, policymakers and researchers involved in performance assessment. The main policy implications of this thesis are:

- Narrative comments are indispensable for the educational value of MSF and should be included in all questionnaires. Respondents in MSF procedures should be better informed about their task and how to give feedback.

- Mentors should be trained to deliver feedback and stimulate reflection. Important strategies to be taught are: contrasting and collating information, posing reflective questions and goal setting.
- Hospitals need to be informed about the anonymous results of the MSF assessments to encourage a joint responsibility for optimal clinical performance.
- Specialist groups and hospitals should be aware of the positive effect of collegial, reflective learning and are recommended to invest in group reflections.

As for researchers, future studies should further explore the validity of MSF assessments by investigating the relation between the assessments and narrative comments. A second important area for further study is the impact of performance assessment on future practice. This might be studied using long-term interventionist, experimental models.

Samenvatting

In **Hoofdstuk 1** wordt de groeiende aandacht voor het evalueren van het functioneren van individuele artsen nader beschreven. Deze groeiende interesse is een logische ontwikkeling omdat de eisen die vandaag de dag aan artsen gesteld worden veel verder gaan dan vakkennis. Artsen hebben te maken met goedgeïnformeerde patiënten, grote interdisciplinaire teams en een toenemende stroom van medische informatie, toe te passen in de medische praktijk. Dit vraagt om goede communicatieve- en samenwerkingsvaardigheden, empathisch en reflectief vermogen en organisatorische vaardigheden. Vaardigheden die bij uitstek in de praktijk moeten worden geëvalueerd en verbeterd. Literatuuronderzoek laat zien hoe in landen als de Verenigde Staten, Canada en het Verenigd Koninkrijk met deze vraag wordt omgegaan. Hier wordt het functioneren van veel artsen structureel geëvalueerd. Wat echter ontbreekt zijn empirische gegevens over hoe deze evaluaties bijdragen aan beter functioneren en een betere kwaliteit van zorg. Het doel van dit proefschrift is dan ook te onderzoeken hoe een systeem voor het evalueren van individueel functioneren van artsen ontworpen en geïmplementeerd kan worden, wat valide, betrouwbaar, bruikbaar en effectief is en een bijdrage levert aan het beter functioneren van artsen.

Hoofdstuk 2: Welke methoden en instrumenten zijn beschikbaar om het functioneren van individuele artsen te evalueren in de dagelijkse praktijk en wat is bekend over de praktische toepasbaarheid, validiteit, betrouwbaarheid en impact op de praktijk?

In hoofdstuk twee wordt een literatuuronderzoek beschreven naar de beschikbare methoden en instrumenten om het functioneren van artsen te evalueren. 64 artikelen voldeden aan de inclusiecriteria. In deze artikelen kon een breed scala aan methoden en instrumenten worden geïdentificeerd. Zes daarvan kunnen worden gebruikt om individueel functioneren te evalueren, namelijk: video-observatie, simulatiepatiënten, directe observatie, medisch dossier onderzoek, 360-graden feedback en portfoliobeoordeling. De praktische toepasbaarheid, validiteit, betrouwbaarheid en educatieve impact van de verschillende methoden loopt echter uiteen. Video-observatie, directe observatie en simulatiepatiënten vraagt een grotere tijdsinvestering. Gemiddeld 2 tot 3 uur per persoon zijn nodig om tot een betrouwbaar oordeel te komen. 360-graden feedback blijkt de methode die het best praktisch toepasbaar is qua tijd. Betrouwbare resultaten worden hiermee bereikt met 1 uur tijdsinvestering. De review laat ook zien dat de empirische onderbouwing van de validiteit van veel 360-graden feedbackinstrumenten onvoldoende is. Enkele studies voldeden niet aan de huidige standaarden van instrument ontwikkeling en de onderbouwing van de constructvaliditeit en criterionvaliditeit ontbrak vaak. Onderzoek naar het effect van het evalueren van artsen staat nog in de kinderschoenen en is vaak gebaseerd op zelfrapportage. Uit ons literatuuronderzoek blijkt dat de meerderheid van de artsen die een evaluatie met 360-graden feedback of portfolio ondergaat tevreden is met de evaluatie en ook van mening is dat zijn/ haar functioneren is verbeterd (tussen de 61 en 72 procent in verschillende onderzoeken). Het is nog onduidelijk wat het effect is van het evalueren van individuele artsen op de kwaliteit van zorg.

Hoofdstuk 3: Wat zijn de psychometrische eigenschappen van drie nieuwe 360-graden feedback instrumenten, die zijn aangepast voor de Nederlandse situatie en hoeveel evaluaties zijn nodig per specialist voor een betrouwbare evaluatie?

360-graden feedback kan pas van waarde zijn wanneer artsen, managers en patiënten vertrouwen hebben in de validiteit en betrouwbaarheid van de instrumenten die worden gebruikt. Voor dit hoofdstuk werden drie aspecten van validiteit en betrouwbaarheid onderzocht, namelijk: (1) de psychometrische eigenschappen van de instrumenten gebruikt in de 360-graden feedback, (2) de relatie tussen de evaluaties vanuit de verschillende perspectieven van collega-specialist, medewerker, patiënt en de zelfevaluatie, (3) het aantal evaluaties dat nodig is voor een betrouwbaar oordeel. Er deden 150 specialisten uit 26 ziekenhuizen aan dit onderzoek mee. Zij werden geëvalueerd door 864 collega-artsen, 894 medewerkers (verpleegkundigen, OK-personeel, polikliniekassistentes) en 1960 patiënten. Principiële componenten analyse laat zien dat er zes onderliggende factoren zijn in het instrument voor collega-artsen die in totaal 67 procent van de variantie verklaren. Deze schalen zijn: samenwerking en zelfinzicht, medisch inhoudelijk handelen, coördinatie en continuïteit van zorg, spoedeisende zorg en time management en verantwoordelijkheid. Het instrument voor medewerkers bestaat uit drie onderliggende factoren, namelijk samenwerking met andere zorgverleners, communicatie met patiënten en patiëntenzorg. Samen verklaren deze factoren 70 procent van de variantie. Principiële component analyse voor de evaluaties van patiënten laat één onderliggende schaal zien, namelijk patiëntgerichtheid die 60 procent van de variantie in scores verklaard. Alle instrumenten blijken intern consistent met Cronbach's alpha's groter dan 0.94. Slechts 2 procent van de evaluaties wordt beïnvloed door bias: leden uit dezelfde maatschap scoren hun collega's iets hoger dan niet-maatschapsleden. Andere factoren zoals geslacht en leeftijd en de lengte van de samenwerkingsrelatie zijn niet van invloed. We vonden wel een klein maar significant verschil tussen mannelijke en vrouwelijk medewerkers: mannelijke medewerkers geven een significant lager oordeel in vergelijking met vrouwelijke medewerkers. Daarnaast zijn de evaluaties van collega-artsen, medewerkers en patiënten over één specialist significant met elkaar gecorreleerd. (Pearsons r= 0.210-0.352, p-waarde <0.001) Echter geen van deze evaluaties is gecorreleerd met de zelfevaluatie. In tegenstelling tot eerdere studies die lieten zien dat acht collega's, acht medewerkers en 25 patiënten nodig zijn, concluderen wij dat vijf collega-evaluaties, vijf medewerker evaluaties en 11 patiëntevaluaties voldoende zijn om betrouwbare resultaten te bereiken.

Hoofdstuk 4: Wat is de praktisch toepasbaarheid en effectiviteit van een evaluatiesysteem gebaseerd op 360-graden feedback en portfolio in acht Nederlandse algemene ziekenhuizen en welke onderwerpen worden besproken in de evaluaties?

Op basis van de resultaten uit hoofdstuk twee selecteerden we twee methoden voor het evaluatiesysteem voor medisch specialisten: 360-graden feedback en een portfolio. We evalueerden de implementatie van dit systeem. Het evaluatiesysteem bestaat uit: (i) een van de

drie verschillende 360-graden feedback instrumenten (Violato's Physician Achievement Review (PAR), Ramsey's ABIM of de Nederlands methodiek van Appraisal en Assessment, (ii) portfolio, (iii) evaluatiegesprek met een mentor (iv) persoonlijk ontwikkelingsplan. Aan deze pilotstudie namen 109 specialisten en 38 mentoren deel. De evaluatie laat zien dat het gemiddeld 8 uur tijd kost om één specialist te beoordelen. Dit bestaat uit: voorbereidingstijd van de mentor (2½ uur), voorbereidingstijd van de specialist (2½ uur), tijdsduur van het evaluatiegesprek (1 uur) en de tijd van de respondenten om de 360-graden feedback in te vullen (2 uur). Communicatie, samenwerking en organisatie zijn de onderwerpen die het meest worden besproken in het evaluatiegesprek. Dit is onderwerp van discussie in respectievelijk 78 procent, 74 procent en 71 procent van de gesprekken. Onderwerpen gerelateerd aan andere competenties zoals: gezondheidsbevordering, professionaliteit en medisch inhoudelijk handelen worden minder vaak besproken. Uit de resultaten blijkt dat 89 procent van de specialisten deelname aan de evaluatiegesprekken zou aanbevelen aan collega's. Tweederde van de specialisten geeft aan van plan te zijn zijn/haar functioneren te verbeteren op een of meer gebieden naar aanleiding van de evaluatie. Met een regressie-analyse gebaseerd op generalized estaming equations (GEE) zijn deze resultaten verder geanalyseerd. We zien dat specialisten significant tevredener zijn met 360-graden feedback die open opmerkingen bevat. (regressie coëfficiënt Beta=0.114, p-waarde <0.001) Ook blijkt dat het effect groter wordt ervaren door specialisten als ook medewerkers worden meegenomen in de 360-graden feedback. (regressie coëfficiënt beta =0.115 voor Appraisal en Assessment, regressie coëfficiënt beta = 0.105 voor Violato's PAR , p-waarde<0.001).

Hoofdstuk 5: Wat zijn bevorderende en belemmerende factoren voor artsen om verbeterpunten die naar voren komen met 360-graden feedback te implementeren in de praktijk?

Hoofdstuk 5 exploreert welke factoren stimulerend of belemmerend werken voor het implementeren van verbetersuggesties uit de 360-graden feedback. Een jaar na de eerste evaluatie hebben we hiervoor 23 specialisten geïnterviewd. We selecteerden doelgericht specialisten op verschillen in ziekenhuis, specialisme, werkervaring, geslacht en de mening die ze rapporteerden in eerder onderzoek. We vroegen hen of zij de voorgenomen verbeterplannen hadden uitgevoerd of niet en waarom. De interviews zijn daarom een jaar na de 360-graden feedback beoordeling en het evaluatiegesprek met de mentor uitgevoerd. Twee onderzoekers hebben onafhankelijk van elkaar een score toegekend aan het nivo van verbetering dat de specialisten rapporteerden. Deze nivo's zijn: bewustzijn van een verbeterdoel (1), acceptatie van een verbeterdoel (2), het uitvoeren van een verbeterdoel (3), het behouden van een verbetering (4). Elf specialisten rapporteren een concrete verbetering in de praktijk. De overige twaalf geven aan dat ze geen concrete stappen hebben genomen. Alle specialisten noemen in het interview factoren die volgens hen stimulerend of juist bevorderend werken. Deze konden worden onderverdeeld in vier belangrijke thema's : 1) contextuele factoren gerelateerd aan werkdruk, openheid, sociale steun, ondersteuning vanuit het management, vrije markt principes en publiek wantrouwen; 2) factoren gerelateerd aan feedback; 3) kenmerken van het evaluatiesysteem

zoals mentoren en een portfolio om reflectie te bevorderen, concrete verbeterdoelen en jaarlijkse vervolginterviews; 4) individuele factoren, zoals self-efficacy en motivatie. Specialisten die verbeterden noemen de gespreksvaardigheden van mentoren en het stellen van doelen als stimulerende factoren. We kunnen concluderen dat ondanks belemmerende factoren als hoge werkdruk en de financiering en organisatie van de huidige Nederlandse gezondheidszorg, 360-graden feedback kan leiden tot een verbetering wanneer mentoren artsen begeleiden in het omgaan met de feedback en daarnaast ook (zelf)reflectie stimuleren. Daarnaast laat onze studie echter zien dat ziekenhuizen er vaak niet in slagen om een collegiaal klimaat te creëren wat reflectie ondersteunt.

Hoofdstuk 6: Hoe zien mentoren hun rol in evaluatiegesprekken en wat zijn volgens mentoren effectieve strategieën om feedback te geven en reflectie te stimuleren?

Na ons onderzoek naar de mening van specialisten over 360-graden feedback, zijn we in de mening van mentoren gedoken. Het doel van de studie in hoofdstuk 6 is na te gaan hoe mentoren hun rol ervaren en vervullen bij de 360-graden evaluaties. We hebben hiervoor gebruik gemaakt van een combinatie van kwalitatieve en kwantitatieve onderzoeksmethoden gebaseerd op vragenlijsten en interviews. Analyse van de interviews werd gedaan met hulp van de 'grounded theory'-methode. We hebben alle mentoren gevraagd een vragenlijst in te vullen. 89 procent van de mentoren geeft aan graag door te gaan als mentor. Daarnaast laten de interviews zien dat mentoren verschillende strategieën gebruiken om effectief feedback te geven, waaronder: het samenvatten en tegen elkaar afzetten van informatie, vragen stellen die reflectie stimuleren en doelen stellen. Mentoren zijn unaniem van mening dat objectiviteit cruciaal is voor een goed evaluatiegesprek met een collega-arts. We zien echter ook dat mentoren moeite hebben met de gevraagde tijdsinvestering, het kennen van de arts met wie je het gesprek voert en het beheersen van specifieke gespreksvaardigheden. Een andere interessante bevinding is dat mentoren zelf ook positieve effecten opmerken van het voeren van evaluatiegesprekken met collega-artsen, namelijk: meer wederzijds respect en saamhorigheid en het verbeteren van hun gespreksvaardigheden.

Hoofdstuk 7: Welke factor(en) hebben de grootste invloed op de veranderingen die specialisten rapporteren in hun functioneren als gevolg van de 360-graden feedback?

Hoofdstuk 7 beschrijft een observationele studie die gebaseerd is op vragenlijstonderzoek onder 236 medisch specialisten in 26 Nederlandse ziekenhuizen. Deze studie bouwt voort op de bevindingen in hoofdstuk waarin we vonden dat verschillende contextuele factoren het gebruik van 360-graden feedback en een verbetering in de praktijk beïnvloeden. Om deze studie optimaal te onderbouwen werd literatuur uit andere vakgebieden (psychologie en human resources) onderzocht. Hieruit bleek dat factoren zoals leeftijd, geslacht en de discrepantie tussen de zelfevaluatie en de evaluatie door anderen, het gebruik van 360-graden feedback kunnen beïnvloeden. In dit hoofdstuk onderzoeken wij of deze bevindingen ook voor artsen gelden.

Specialisten werden verzocht om een half jaar na het evaluatiegesprek een korte vragenlijst in te vullen over hun tevredenheid en de mate waarin ze hun functioneren hadden verbeterd naar aanleiding van de evaluaties. Het responspercentage was 52 procent. De afhankelijke variabele was zelfrapportage van verandering als gevolg van de 360-graden feedback. Het verband tussen de drie categorieën onafhankelijke variabelen (persoonlijke kenmerken zoals leeftijd en geslacht, tevredenheid met de evaluaties in het algemeen en de ontvangen oordelen op de 360-graden feedback) en de gerapporteerde verandering in gedrag werd onderzocht met multivariate regressie-analyse. Een kleine meerderheid (55 procent) van de specialisten geeft aan een of meer aspecten van zijn/ haar functioneren te hebben verbeterd. Twee variabelen zijn hierop significant van invloed, namelijk: de ervaren kwaliteit van de mentor (regressie coëfficiënt = 0.527, $p< 0.05$) en de scores die men krijgt van collega's (regressie coëfficiënt Beta=-0.157, $p<0.05$). De scores die men krijgt van collega's blijken negatief geassocieerd met de gerapporteerde verandering. Dit betekent dat specialisten aangeven meer te veranderen naarmate ze lagere scores krijgen van collega's. Deze twee variabelen verklaren samen 34 procent van de variantie. Daarnaast blijkt dat de gemiddelde zelfevaluaties van vrouwen significant lager zijn dan die van mannen. Dit onderzoek brengt aan het licht dat de kwaliteit van de mentoren en de scores gegeven door collega's de belangrijkste stimulerende factoren zijn voor een verbetering van het functioneren naar aanleiding van 360-graden feedback.

In **hoofdstuk 8,** de discussie, worden de belangrijkste resultaten van dit proefschrift besproken en gerelateerd aan andere studies en eerdere onderzoeken. Het literatuuronderzoek bracht een uitgebreid arsenaal aan methoden en instrumenten om het individueel functioneren van artsen te evalueren aan het licht. 360-graden feedback bleek het best praktisch toepasbaar in de dagelijkse praktijk. Specifiek voor Nederland hebben we de validiteit en betrouwbaarheid van 360-graden feedback vragenlijsten kunnen onderbouwen. De studies in dit proefschrift laten zien dat verschillende factoren de verandering beïnvloeden die door artsen wordt ingezet na 360-graden feedback. Het opnemen van feedback door medewerkers, open opmerkingen en een mentor stimuleren het gebruik van 360-graden feedback zoals ook uit eerdere onderzoeken in andere settings bleek. Daarnaast hebben we factoren ontdekt die niet in eerder onderzoek werden gevonden zoals contextuele factoren gerelateerd aan de maatschappij, gezondheidszorg systeem en de maatschap en individuele factoren zoals self-efficacy en motivatie. In dit hoofdstuk presenteren wij een theoretisch kader wat de verschillende factoren bevat die de impact van evaluatiegesprekken van artsen beïnvloeden. Dit theoretisch kader kan worden gebruikt om richting te geven aan toekomstige studies. Het hoofdstuk sluit af met een samenvatting van de implicaties voor artsen, mentoren, beleidsmakers en onderzoekers.

De belangrijkste implicaties voor de praktijk van dit proefschrift zijn:

- Open opmerkingen zijn onmisbaar voor de educatieve waarde van 360-graden feedback en zouden onderdeel moeten zijn van alle vragenlijsten. Respondenten in 360-graden

feedback procedures moeten beter geïnformeerd worden over het belang van open opmerkingen en hoe ze deze moeten geven.

- Mentoren moeten worden getraind in het geven van 360-graden feedback en het stimuleren van reflectie. Belangrijke strategieën hierbij zijn: het samenvatten en tegen elkaar afzetten van verschillende informatie, het stellen van vragen die aanzetten tot reflectie en stimuleren tot doelen stellen.
- Ziekenhuizen moeten op de hoogte worden gebracht van de resultaten van de evaluatiegesprekken – al dan niet anoniem - zodat een gezamenlijke verantwoordelijkheid ontstaat voor optimaal functioneren van artsen.
- Maatschappen en ziekenhuizen moeten zich meer bewust worden van het positief effect van reflectief leren en zouden moeten investeren in groepsreflecties.

Voor onderzoekers biedt dit proefschrift meerdere aanknopingspunten voor vervolgonderzoek. De validiteit van 360-graden beoordelingen zou beter kunnen worden onderbouwd door de relatie met open opmerkingen na te gaan. Een tweede belangrijk onderzoeksterrein betreft de impact van het evalueren van artsen op hun toekomstig functioneren. Dit kan worden onderzocht met langlopend onderzoek en gebruikmakend van experimentele modellen.

Dankwoord

With good company, no road is long.

Beste Richard, ik schrok soms wel even als ik je gedetailleerde commentaar op een artikel zag. Je wist altijd de vinger op de zere plek te leggen. Ik bewonder je om je vele kwaliteiten; van goed onderzoek doen, promovendi stimuleren en bijsturen tot het bedrijven van politiek. Je liet me steeds meer mijn eigen weg bewandelen en ontdekken hoe ik het aan moest pakken. Bedankt voor alles wat ik van je heb mogen leren.

Beste Hub, ik waardeer je pragmatische aanpak en betrokkenheid waarbij je altijd wist wat er ook politiek speelde en waarmee ik in de knoop zat. Je wist ook goed de link te leggen met andere projecten die er spelen. Bedankt voor de vrijheid die je me gaf om me te ontplooien.

Lieve Kiki, ik waardeer je op vele fronten: je doorzettingsvermogen, oog voor detail, persoonlijke betrokkenheid en jouw gave om een team te vormen met een duidelijk missie. Je bent een rolmodel voor me geweest op meerdere fronten. Dankjewel voor de fijne gesprekken, en de inspiratie die je me gaf om een weg te bewandelen naar meer rust in mijn hoofd. Inmiddels kom je ook weleens zomaar bij mij over de vloer en ik hoop dat dat zo blijft.

Beste Onyi, I have great respect for the way you practice research and teach most difficult statistics. You took care that I started with a detailed proposal, stimulated me to work on its theoretical foundations. I will never forget you stimulated me to submit an article 3 months after I started my PhD to JAMA's special issue. The best advice you could have given me!

Vanuit de Orde van Medisch Specialisten zijn Ko van de Klundert, Dieudonné Trip en Marjon Kallewaard bij dit onderzoek betrokken geweest. Dank voor de goede samenwerking. De uitvoering van de onderzoeksprojecten in dit proefschrift had nooit plaats kunnen vinden zonder de deelname van enthousiaste specialisten en ziekenhuizen die een voortrekkersrol vervulden. Daarnaast ben ik de begeleidingscommissie IFMS van de Orde van Medisch Specialisten dankbaar voor het meedenken over het project IFMS en het daaraan gekoppelde onderzoek.

Toen ik zes jaar geleden aan dit project begon, had ik eerlijk gezegd nooit durven dromen dat het af zou komen met de huisartsopleiding ernaast, waar ook een groot deel van mijn hart ligt. Toch is het gebeurd en dat heb ik zeker ook aan de bemoedigende woorden, gezellige avondjes en steun van anderen te danken. Zonder iemand tekort te willen doen wil ik toch een paar van jullie noemen.

Dankjewel aan mijn onderzoeksassistenten Geertje van de Ven en Juliette Cruijsberg voor alle analyses en meedenkwerk. Ik ben jullie beiden veel dank verschuldigd voor het goede werk dat jullie hebben verricht en de gezellige samenwerking.

Alle medewerkers van IQ healthcare wil ik bijzonder bedanken voor de vele gezellige momenten die ik heb mogen beleven en de inspiratie die jullie me gaven op allerlei vlakken. In het bijzonder wil ik Jolanda van Haren noemen. Ongelofelijk hoe netjes jij de lay-out hebt gedaan. Ik schoot al in de stress bij enkel de

gedachte dat een pdf artikel weer in Word moest worden omgezet. Maar voor jou is dat een fluitje van een cent. Jeannette en Myriam wil ik bedanken voor al het geregel in de afgelopen jaren, het maken van afspraken, versturen van vragenlijsten etc. Anja, bedankt voor je hulp bij alle P&O dingen en financiën. Ook al zo prettig gestructureerd. Dankzij jou ging er nooit iets fout met alle declaraties. Marjan Faber, bedankt voor de razend snelle start die je me gaf bij de review en hulp toen ik net begonnen was.

Lief IQ maatje Linda, dankjewel voor de gezellige kletsavonden waarin we elkaar goed konden relativeren en onze wederzijdse liefde voor onzinnige tijdschriften deed opbloeien. Viva la vida! Not if but when, deze mantra zal ik nooit vergeten.

Lieve Erik, jij leerde mij de kunst van het schrijven en onderzoeken toen ik nog student-assistent bij je was. Inmiddels ben je een erg goede vriend geworden. Na mijn vertrek naar Nijmegen heb je me er op cruciale momenten doorheen gesleept. Niet alleen op onderzoeksgebied heb je een mentorrol in mijn leven vervuld. Ik ben verheugd dat je achter me staat op deze dag.

Beste Mereke, reuze bedankt voor je hulp bij het redigeren van mijn Engelse teksten. Dankzij jouw hulp is mijn Engels zeker verbeterd.

Lieve collega's van praktijk de Grote Rivieren, ik voel me als een vis in het water bij jullie en geniet elke maandag en woensdag met volle teugen. Jullie krijgen nu jullie welverdiende feestje! Barbera en Petra, ik ben blij dat ik jullie beiden als rolmodel en leermeesters ben tegengekomen. Ik heb een hele rugzak vol aan bagage mee gekregen van jullie en ik ben verheugd dat ik de kans heb gekregen om huisarts te zijn bij jullie.

Katinka en Ferry, mijn opleiders in het eerste jaar van de huisartsopleiding in de (UHP). Dankjewel voor de goede start die ik kreeg als huisarts en voor jullie interesse en stimulerende woorden bij de afronding van dit proefschrift.

Lieve intervisie collega's Juul, Kasper, Lisette, Chris en Esther: bedankt voor jullie interesse en de fijne intervisie bijeenkomsten. Ik hoop dat we hiermee nog lang door kunnen gaan.

Lieve Heusden crew. Ik heb genoten van de Heusden weken waarin we naast onderzoeken ook hele boeiende gesprekken hebben gehad over het leven, spiritualiteit en de oliebol en champagnefles van 2010. Renée, bedankt voor je ontzettende stimulans op vrijdag, de dagen dat ik ECHT het proefschrift heb afgeschreven. Ik hoop je in de toekomst nog vaak te zien!

Alle lieve vrienden en vriendinnen op afstand en dichtbij: dank jullie wel dat ik een stukje van mijn leven met jullie mag delen en de leuke momenten die we samen hebben.
In het bijzonder noem ik een paar van jullie:

Lieve Marije, wat hebben we veel gelachen de laatste jaren en gelukkig ook meer nu we beiden onze proefschriften hebben afgerond. Het begon in Maastricht met luchtgitaar spelen en allerhande vreemde toestanden in huisartspraktijken in het heuvelland. Ik ben reuze blij dat we ondanks al onze

omzwervingen nog steeds goede vriendinnen zijn. Ik bewonder je doorzettingsvermogen en je oprechte betrokkenheid bij de mensen om je heen.

Lieve Tanya, onze vriendschap begon op reis met zijn tweeën door Australië. Waar jij me wat meer actie bijbracht en ik jou wat meer rust. Een mooie balans ontstond. Zo'n reis zal er niet meer inzitten maar wat mij betreft nog wel heel veel jaren trouwe vriendschap!

Lieve Lydia, al heel lang loop je mee in mijn leven. Heel blij ben ik daarom. Soms een tijdje niet, dan weer heel intensief als er wat gebeurt. Je bent 'unne goeie'. Bedankt voor de gezelligheid en je steun als ik die nodig had.

Lieve Jasper en Maartje. Tja, 33 jaar na 1978 hebben Jasper en Kristof nog altijd contact en nu in Amsterdam. Dank jullie wel voor de leuke afleidende momenten. Ik verheug me op nog vele mooie tochten op het water en weekendjes Ardennen. Jasper, reuze bedankt voor je hulp bij de kaft.

Lieve Tako en Jeff, met jullie ben je altijd verzekerd van een leuke avond. Tako, bedankt dat je je liefde voor herten en natuur met me wilt delen. Ik kijk uit naar onze volgende herten-bronst-strooptocht of schaatstocht met chocomel-baileys.

Lieve Nijmegen crew: Willem, Judith, Teun, Carolien, Peter, Femke en Jasper. Wat een massaproductie aan proefschriften produceren we. Er zullen er nog 2 volgen! Dat we nog maar veel leuke feestjes met elkaar mogen vieren. Carolien en Teun, wat een fijne gesprekken hebben we toch altijd. Jullie zijn heel belangrijk voor ons geworden in de laatste jaren. Lieve Carolien, ik ben blij met een vriendin als jij en hoop dat we nog maar veel gesprekken mogen hebben over het combineren van ambitie met een leuk leven!

Lieve Joset, ook jij hoort zeker thuis in dit dankwoord. Ik heb veel respect voor jouw doorzettingsvermogen en ik bewonder je rechtvaardigheidsgevoel! Dankjewel voor je interesse en steun.

Lieve Francoise, Ellen, Inge en Marc. Ik heb het getroffen met jullie als schoonfamilie! Jullie betrokkenheid was er op vele momenten. Ik ben er trots op dat we er zo voor elkaar zijn geweest in de laatste 3 jaar. Lieve Mari, wat hadden we je er graag bij gehad. En bij nog heel veel andere momenten die er dagelijks zijn en nog zullen komen. In gedachten ben je bij ons.
Lieve Oma Franse, we zijn blij dat je er bij bent op deze dag.

Lieve Joost en Marga. Jullie zijn de meest fantastische ouders die ik me had kunnen wensen. Niet alleen vanwege de reuze gezellige jeugd en basis die jullie me gaven. Maar vooral ook door de zelfstandigheid en de mentaliteit van hard werken en klaar staan voor een ander die ik met de paplepel kreeg ingegoten. Heel mijn leven zal ik daarvan profiteren. Dankjewel voor jullie jarenlange liefde, steun en stimulans waar ik altijd op kon rekenen. Daarom draag ik dit proefschrift aan jullie op.

Lieve Jorn, Feyenoord-fan in hart en nieren. Ik weet nu ook wat dagenlang cijferwerk is. Veel respect heb ik voor je werk en je relativeringsvermogen. Ik hoop dat je in de toekomst meer in Nederland zult zijn en we kunnen blijven genieten van veel gezelligheid.

Lieve Rens, dankjewel dat je mijn paranimf wilt zijn. Wat houdt jij me heerlijk op de grond! Als je een paper moet schrijven van 5 pagina's, sms je me 'lief zusje, het zijn maar 5 pagina's, ik moet nu alleen nog gaan voor kwaliteit'. Rens, je zult je leven kunnen profiteren van je relativeringsvermogen en je rotsvaste vertrouwen in jezelf en de mensheid om je heen.

Liefste Kristof, het was tijdens een wandeling in Toscane. Jij weet vast niet meer waar het over ging. Mij staat ons gesprek nog altijd helder voor de geest. Ik schrok een beetje van je felheid maar je wilde me vooral duidelijk maken dat onderzoek doen niet samen gaat met slaafs zijn. Een heerlijke metafoor zoals alleen jij die kunt verkondigen! Nu kan ik zeggen dat je vanaf toen -en niet alleen op dat moment- veel meer hebt betekend dan je zelf realiseert. Je daarvoor bedanken zou afbreuk doen aan onze liefde. Want die is veel meer dan de steun die je me geeft en niet in honderd proefschriften uit te leggen. Nou ja dan toch een poging: je haalt het mooiste bij mij naar boven. Mi piace solo tu.

Karlijn

Curriculum Vitae

Karlijn Overeem was born on January 22nd 1980 as the eldest in a family with three children. She grew up in Tilburg where she obtained her pre university diploma in 1998 from the Pauluslyceum. After that she studied medicine at Maastricht University, performing internships in Sweden and Australia. In 2005 she obtained her medical degree with honors. During her studies Karlijn was active as a student member of the faculty board and student assistant in a portfolio project for medical students. Since 2006 Karlijn worked at UMC St Radboud's research center: IQ healthcare, where she combined her PhD research into the Individual Performance of Medical Specialists with her specialization to become a general practitioner. For this she obtained AGIKO-funds from ZonMW. In 2008, she won the Patil award for the best presentation in the category 'Research in Medical Education' at the AMEE-conference in Prague. Currently she is active as a general practitioner at the Grote Rivieren practice in Amsterdam. The city in which she lives since 2007 together with her partner Kristof Franse (1978).

Karlijn Overeem werd geboren op 22 januari 1980 als oudste van een gezin van drie. Ze groeide op in Tilburg en behaalde daar in 1998 het V.W.O diploma aan het Pauluslyceum. Daarna studeerde ze geneeskunde aan de Universiteit Maastricht. Ze liep tijdens haar studie stage in Zweden en Australië. In 2005 behaalde ze cum laude haar artsexamen. Naast haar geneeskunde-opleiding was Karlijn student-lid van het faculteitsbestuur en student assistent op een portfolioproject voor geneeskundestudenten. Sinds 2006 is Karlijn als onderzoeker verbonden aan de afdeling IQ healthcare van het UMC St Radboud. Ze combineerde haar promotietraject naar het Individueel Functioneren van Medisch Specialisten met de huisartsopleiding van het VU Medisch Centrum in Amsterdam en verkreeg hiervoor een AGIKO-subsidie van ZonMW. In 2008 won ze de prijs voor de beste presentatie in de categorie 'Research in Medical Education' bij de AMEE-conferentie in Praag. Momenteel is ze werkzaam als huisarts in praktijk de Grote Rivieren in de Amsterdamse Rivierenbuurt. Sinds 2007 woont ze samen met haar partner Kristof Franse (1978) in Amsterdam.